Safe Probability

Peter Grünwald

April 8, 2016

Abstract

We formalize the idea of probability distributions that lead to reliable predictions about some, but not all aspects of a domain. The resulting notion of 'safety' provides a fresh perspective on foundational issues in statistics, providing a middle ground between imprecise probability and multiple-prior models on the one hand and strictly Bayesian approaches on the other. It also allows us to formalize fiducial distributions in terms of the set of random variables that they can safely predict, thus taking some of the sting out of the fiducial idea. By restricting probabilistic inference to safe uses, one also automatically avoids paradoxes such as the Monty Hall problem. Safety comes in a variety of degrees, such as 'validity' (the strongest notion), 'calibration', 'confidence safety' and 'unbiasedness' (almost the weakest notion).

1 Introduction

We formalize the idea of probability distributions that lead to reliable predictions about some, but not all aspects of a domain. Very broadly speaking, we call a distribution $\tilde{P}$ *safe* for predicting random variable U given random variable V if predictions concerning U based on $\tilde{P}(U|V)$ tend to be as good as one would expect them to be if $\tilde{P}$ were an accurate description of one's uncertainty, even if $\tilde{P}$ may not represent one's actual beliefs, let alone the truth. Our formalization of this notion of 'safety' has repercussions for the foundations of statistics, providing a joint perspective on issues hitherto viewed as distinct:

1. All models are wrong...[1] Some statistical models are evidently both entirely wrong yet very useful. For example, in some highly successful applications of Bayesian statistics, such as latent Dirichlet allocation for topic modeling (Blei et al., 2003), one assumes that natural language text is i.i.d., which is fine for the task at hand (topic modeling) — yet no-one would want to use these models for predicting the next word of a text given the past. Yet, one can use a Bayesian posterior to make such predictions any way — Bayesian inference has no mechanism to distinguish between 'safe' and 'unsafe' inferences. Safe probability allows us to impose such a distinction.

2. The Eternal Discussion[2] More generally, representing uncertainty by a single distribution, as is standard in Bayesian inference, implies a willingness to make definite predictions about random variables that, some claim, one really knows nothing about. Disagreement on

[1] ...yet some are useful, as famously remarked by Box (1979).

[2] When the single-vs. multiple-prior issue came up in a discussion on the *decision-theory forum* mailing list, the well-known economist I. Gilboa referred to it as 'the eternal discussion'.

this issue goes back at least to Keynes (1921) and Ramsey (1931), has led many economists to sympathize with *multiple-prior models* (Gilboa and Schmeidler, 1989) and some statisticians to embrace the related *imprecise probability* (Walley, 1991, Augustin et al., 2014) in which so-called 'Knightian' uncertainty is modeled by a *set* $\mathcal{P}^*$ of distributions. But imprecise probability is not without problems of its own, an important one being *dilation* (Example 1 below). Safe probability can be understood as starting from a set $\mathcal{P}^*$, but then *mapping* the set of distributions to a single distribution, where the mapping invoked may depend on the prediction task at hand — thus avoiding both dilation and overly precise predictions. The use of such mappings has been advocated before, under the name *pignistic transformation* (Smets, 1989, Hampel, 2001), but a general theory for constructing and evaluating them has been lacking (see also Section 5).

3. Fisher's Biggest Blunder[3] Fisher (1930) introduced *fiducial inference*, a method to come up with a 'posterior' $\tilde{P}(\theta \mid X^n)$ on a model's parameter space based on data X^n, but without anything like a 'prior', in an approach to statistics that was neither Bayesian nor frequentist. The approach turned out problematic however, and, despite progress on related *structural inference* (Fraser, 1968, 1979) was largely abandoned. Recently, however, fiducial distributions have made a comeback (Hannig, 2009, Taraldsen and Lindqvist, 2013, Martin and Liu, 2013, Veronese and Melilli, 2015), in some instances with a more modest, frequentist interpretation as *confidence distributions* (Schweder and Hjort, 2002, 2016). As noted by Xie and Singh (2013), these 'contain a wealth of information for inference', e.g. to determine valid confidence intervals and unbiased estimation of the median, but their interpretation remains difficult, viz. the insistence by Hampel (2006), Xie and Singh (2013) and many others that, although $\tilde{P}(\cdot \mid X^n)$ is defined as a distribution on the parameter space, the parameter itself is not random. Safe probability offers an alternative perspective, where the insistence that 'θ is not random' is replaced by the weaker (and perhaps liberating) statement that 'we can treat θ as random' *as long as we restrict ourselves to safe inferences about it*' — in Section 3.1 we determine precisely what these safe inferences are and how they fit into a general hierarchy:

4. The Hierarchy Pursuing the idea that some distributions are reliable for a smaller subset of random variables/prediction tasks than others, leads to a natural *hierarchy* of safeties — a first taste of which is in Figure 1 on page 5, with notations explained later. At the top are distributions that are fully reliable for whatever task one has in mind; at the bottom those that are reliable only for a single task in a weak, average sense. In between there is a natural place for distributions that are *calibrated* (Example 2 below), that are *confidence–safe* (i.e. valid confidence distributions) and that are *optimal for squared-error prediction*.

5. "The concept of a conditional probability with regard to an isolated hypothesis..."[4] Upon first hearing of the Monty Hall (quiz master, three doors) problem (vos Savant, 1990, Gill, 2011), most people naively think that the probability of winning the car is the same whether one switches doors or not. Most can eventually, after much arguing, be

[3] While Fisher is generally regarded as (one of) the greatest statisticians of all time, fiducial inference is often considered to be his 'big blunder' — see Hampel (2006) and Efron (1996), who writes *Maybe Fisher's biggest blunder will become a big hit in the 21st century!*

[4] ... whose probability equals 0 is inadmissible," as remarked by Kolmogorov (1933). As will be seen, safe probability suggests an even more radical statement related to the Monty Hall sanity check.

convinced that this is wrong, but wouldn't it be nice to have a simple sanity check that *immediately* tells you that the naive answer must be wrong, without even pondering the 'right' way to approach the problem? Safe probability provides such a check: one can immediately tell that the naive answer is *not safe*, and thus cannot be right. Such a check is applicable more generally, whenever conditioning on events rather than on random variables (Example 4 and Section 4).

6. "Could Neyman, Jeffreys and Fisher have agreed on testing?"[5] Ryabko and Monarev (2003) shows that sequences of 0s and 1s produced by standard random number generators can be substantially compressed by standard data compression algorithms such as `rar` or `zip`. While this is clear evidence that such sequences are not random, this method is neither a valid Neyman-Pearson hypothesis test nor a valid Bayesian test (in the tradition of Jeffreys). The reason is that both these standard paradigms require the existence of an *alternative statistical model*, and start out by the assumption that, if the null model (i.i.d. Bernoulli (1/2)) is incorrect, then the alternative must be correct. However, there is no clear sense in which `zip` could be 'correct' — see Section 5. There is a third testing paradigm, due to Fisher, which does view testing as accumulating evidence against h_0, and not necessarily as confirming some precisely specified h_1. Yet Fisher's paradigm is not without serious problems either — see Section 5.

Berger et al. (1994) started a line of work culminating in Berger (2003), who presents tests that have interpretations in all three paradigms and that avoid some of the problems of their original implementations. However, it is essentially an objective Bayes approach and thus inevitably, strong evidence against h_0 implies a high posterior probability that h_1 is true. If one is really doing Fisherian testing, this is unwanted. Using the idea of safety, we can extend Berger's paradigm by stipulating the inferences for which we think it is safe: roughly speaking, if we are in a Fisherian set-up, then we declare all inferences conditional on h_1 to be unsafe, and inferences conditional on h_0 to be safe; if we really believe that h_1 may represent the state of the world, we can declare inferences conditional on h_1 to be safe. But much more is possible using safe probability — a DM can decide, on a case by case basis, what inferences based on her tests would be safe, and under what situations the test results itself are safe — for example, some tests remain safe under optional stopping, whereas others (even Bayesian ones!) do not. While we will report on this application of safety (which comprises a long paper in itself) elsewhere, we will briefly return to it in the conclusion.

7. Further Applications: Objective Bayes, Epistemic Probability Apart from the applications above, the results in this paper suggest that safe probability be used to formalize the status of default priors in *objective Bayesian* inferences, and to enable an alternative look at *epistemic probability*. But this remains a topic for future work, to which we briefly return at the end of the paper.

The Dream Imagine a world in which one would require any statistical analysis — whether it be testing, prediction, regression, density estimation or anything else — to be accompanied by a *safety statement*. Such a statement should list what inferences, the analysists think, can be safely made based on the conclusion of the analysis, and in what formal 'safety' sense. Is the alternative h_1 really true even though h_0 is found to be false? Is the suggested predictive

[5] ...", as asked by Jim Berger (2003).

distribution valid or merely calibrated? Is the posterior really just good for making predictions via the predictive distribution, or is it confidence-safe, or is it generally safe? Does the inferred regression function only work well on covariates drawn randomly from the same distribution, or also under covariate shift? (an application of safety we did not address here but which we can easily incorporate). The present, initial formulation of safe probability is too complicated to have any realistic hopes for a practice like this to emerge, but I can't help hoping that the ideas can be simplified substantially, and a safer practice of statistics might emerge.

Starting with Grünwald (1999), my own work — often in collaboration with J. Halpern — has regularly used the idea of 'safety', for example in the context of Maximum Entropy inference (Grünwald, 2000), and also dilation (Grünwald and Halpern, 2004), calibration (Grünwald and Halpern, 2011), and probability puzzles like Monty Hall (Grünwald and Halpern, 2003, Grünwald, 2013). However, the insights of earlier papers were very partial and scattered, and the present paper presents for the first time a general formalism, definitions and a hierarchy. It is also the first one to make a connection to confidence distributions and pivots.

1.1 Informal Overview

Below we explain the basic ideas using three recurring examples. We assume that we are given a set of distributions $\mathcal{P}^*$ on some space of outcomes $\mathcal{Z}$. Under a frequentist interpretation, $\mathcal{P}^*$ is the set of distributions that we regard as 'potentially true'; under a subjectivist interpretation, it is the *credal set* that describes our uncertainty or 'beliefs'; all developments below work under both interpretations.

All probability distributions mentioned below are either an element of $\mathcal{P}^*$, or they are a *pragmatic distribution* $\tilde{P}$, which some decision-maker (DM) uses to predict the outcomes of some variable U given the value of some other variable V,where both U and V are random quantities defined on $\mathcal{Z}$. $\tilde{P}$ is also used to estimate the quality of such predictions. $\tilde{P}$ (which may be, but is not always in $\mathcal{P}^*$) is 'pragmatic' because we assume from the outset that some element of $\mathcal{P}^*$ might actually lead to better predictions — we just do not know which one.

Example 1 **[Dilation]** A DM has to make a prediction or decision about random variable $U \in \mathcal{U} = \{0, 1\}$ given the value of $V \in \mathcal{V} = \{0, 1\}$. She knows that the marginal probability $P(U = 1) = 0.9$; she suspects that U may depend on V, but has no idea whether U and V are positively or negatively correlated or how strong the correlation is. She may thus model her uncertainty as the set $\mathcal{P}^*$ of *all* distributions P on $\mathcal{Z} = \mathcal{U} \times \mathcal{V}$ that satisfy

$$P(U = 1) = \sum_{v \in \mathcal{V}} P(U = 1, V = v) = 0.9. \tag{1}$$

Given that $V = 1$, what should she predict for U? A standard answer in imprecise probability (Walley, 1991) is to pointwise condition the set $\mathcal{P}^*$, leading one to adopt the probabilities $\mathcal{P}^*(U = 1 \mid V = 1) := \{P(U = 1 \mid V = 1) : P \in \mathcal{P}^*\}$. But this set contains *every* distribution on U, including $P(U = 1 \mid V = 1) = 0$ (the latter would obtain for the $P \in \mathcal{P}^*$ with $P(U = |1 - V|) = 1$). It therefore seems that, after observing $V = 1$, the DM has lost rather than gained information. By symmetry, the same happens after observing $V = 0$, so *whatever DM observes, she loses information* — a phenomenon known as *dilation* (Seidenfeld and Wasserman, 1993). This is intuitively disturbing, and it may perhaps be better to simply

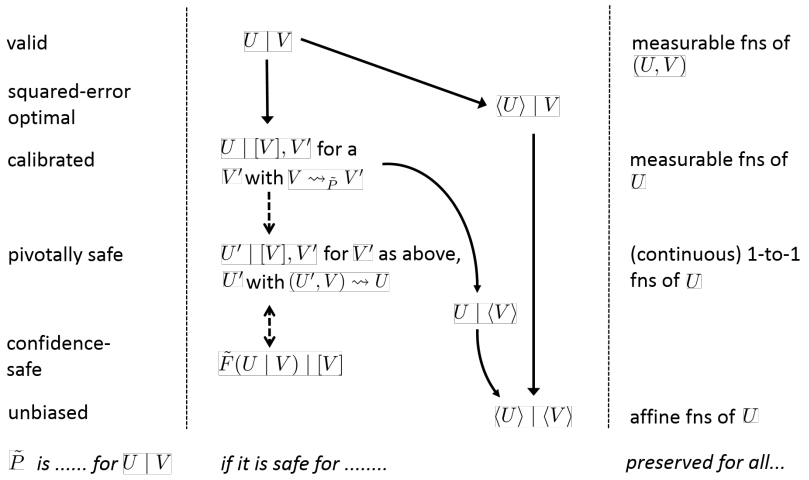

Figure 1: A Hierarchy of Relations for $\tilde{P}$. The concepts on the right correspond (broadly) to existing notions, whose name is given on the left (with the exception of $U \mid \langle V \rangle$, for which no regular name seems to exist). $A \to B$ means that safety of $\tilde{P}$ for A implies safety for B — at least, under some conditions: for all solid arrows, this is proven under the assumption of V with countable range (see underneath Proposition 1). For the dashed arrows, this is proven under additional conditions (see Theorem 2 and subsequent remark). On the right are shown transformations on U under which safety is preserved, e.g. if $\tilde{P}$ is calibrated for $U|V$ then it is also calibrated for $U' \mid V$ for every U' with $U \rightsquigarrow U'$ (see remark underneath Theorem 2). Weakening the conditions for the proofs and providing more detailed interrelations is a major goal for future work, as well as investigating whether the hierarchy has a natural place for *causal* notions, such as $\tilde{P}(U \mid \text{DO}(v))$ as in Pearl's (2009) do-calculus.

ignore V and predict using the distribution that acts as if $U \perp V$ and has

$$\tilde{P}(U = 1 \mid V = v) = P(U = 1) \quad \text{for all } v \in \mathcal{V}, \tag{2}$$

i.e. $\tilde{P}(U = 1 \mid V = v) = 0.9$. While from a purely subjective Bayesian standpoint information is never useless and this seems silly, it is certainly what humans often do in practice, and usually, they get away with it (Dempster, 1968) — for concrete examples see Grünwald and Halpern (2004). Here is where Safe Probability comes in — it tells us that $\tilde{P}$ is *safe* to use, in the following simple sense: for any function $g : \mathcal{U} \to \mathbb{R}$, we have:

$$\text{for all } P \in \mathcal{P}^*, \text{ all } v \in \mathcal{V}: \quad \mathrm{E}_{U \sim P}[g(U)] = \mathrm{E}_{U \sim \tilde{P}}[g(U) \mid V = v]. \tag{3}$$

In particular, if we have a loss function $L : \mathcal{U} \times \mathcal{A} \to \mathbb{R}$ mapping outcomes and actions to associated losses, then, for any action $a \in \mathcal{A}$, we can plug in $g(U) := L(U, a)$ above and then we find that (assuming $\mathcal{P}^*$ contains the truth):

> DM's predictions are guaranteed to be exactly as good, in expectation, as *she would expect them to be if* $\tilde{P}$ *were actually 'true'* — even if $\tilde{P}$ is not true at all.

We immediately add though that if we had a loss function $L' : \mathcal{U} \times \mathcal{V} \times \mathcal{A} \to \mathbb{R}$ which would *itself* depend on V (e.g. if $V = 1$ DM is offered a different bet on U than if $V = 0$) then the $\tilde{P}$ based on ignoring V is not safe any more — (3) may not hold any more, and the actual expectation may be different from DM's. In terms of the formalism we develop below

(Definition 1, 2 and 3), this will be expressed as '$\tilde{P}$ is safe for predicting with loss function L but not loss function L'', or, in formal notation, $\tilde{P}$ is safe for $L(\cdot, a) \mid [V]$ but not for $L'(\cdot, a) \mid [V]$. The intuitive meaning is that DM can safely use $\tilde{P}$ to make predictions against L (her predictions will be as good as she expects) but not against L'. These statements will be immediate consequences of the more general statements '$\tilde{P}$ is safe for $U \mid [V]$ but not safe for $U \mid V$'.

In some cases, we will not be able to come up with a $\tilde{P}$ satisfying (3), and we have to settle for a $\tilde{P}$ that satisfies a weaker notion of safety, such as, for all $P \in \mathcal{P}^*$, all functions g,

$$\mathrm{E}_{V \sim P}\left[\mathrm{E}_{U \sim \tilde{P}}\left[g(U) \mid V\right]\right] = E_{U \sim P}[g(U)], \tag{4}$$

which says that DM predicts as well on average as DM would expect to predict on average if $\tilde{P}$ were true, even though $\tilde{P}$ may not be true. This will be denoted as '$\tilde{P}$ is safe for $U \mid \langle V \rangle$'; and if (4) only holds for g the identity (which makes no difference if $|\mathcal{U}| = 2$, but in general it does) we have the even weaker safety for $\langle U \rangle \mid \langle V \rangle$ (Figure 1). In Section 2.2 we thus obtain five basic notions of safety, varying from weak safety, in an average sense, to very strong safety, safety for $U \mid V$, which essentially means that $\tilde{P}(U \mid V)$ must be the correct conditional distribution.

In this example we used frequentist terminology, such as 'correct' and 'true', and we continue to do so in this paper. Still, a subjective interpretation remains valid in this and future examples as well: if the DM's real beliefs are given by the full set $\mathcal{P}^*$, she can safely act as if her belief is represented by the singleton $\tilde{P}$ as long as she also believes that her loss does not depend on V.

Example 2 **[Calibration]** Consider the weather forecaster on your local television station. Every night the forecaster makes a prediction about whether or not it will rain the next day in the area where you live. She does this by asserting that the probability of rain is p, where $p \in \{0, 0.1, \ldots, 0.9, 1\}$. How should we interpret these probabilities? The usual interpretation is that, in the long run, on those days at which the weather forecaster predict probability p, it will rain approximately $100p\%$ of the time. Thus, for example, among all days for which she predicted 0.1, the fraction of days with rain was close to 0.1. A weather forecaster (DM) with this property is said to be *calibrated* (Dawid, 1982, Foster and Vohra, 1998). Like safety itself, calibration is a *minimal* requirement: for example, a weather forecaster who predicts, each day of the year, that the probability of rain tomorrow is 50% will be approximately calibrated in the Netherlands, but her predictions are not very useful — and it is easily seen that, when using a proper scoring rule, optimal forecasts are calibrated, but calibrated forecasts can be far from optimal. On the other hand, in practice we often see calibrated weather forecasters that predict well, but do not predict with anything close to the 'truth' — their predictions depend on high-dimensional covariates consisting of measurements of air pressure, temperature etc. at numerous locations in the world, and it seems quite unlikely (and, for practical purposes, unnecessary!) that, given any specific values of these covariates, they issue the correct conditional distribution. While calibration is usually defined relative to empirical data, a re-definition in terms of an underlying set of distributions $\mathcal{P}^*$ is straightforward (Vovk et al., 2005, Grünwald and Halpern, 2011), and in Section 2.3 we show that the probabilistic definition of calibration has a natural expression in terms of the safety notions introduced above: $\tilde{P}(U \mid V)$ is calibrated for U if it is safe for $U \mid [V], V'$, for *some* V' with $V \rightsquigarrow V'$ (all notation to be explained) — which implies that (3) is itself an instance of calibration.

Example 3 [Bayesian, Fiducial and Confidence Distributions] We are given a parametric probability model $\mathcal{M} = \{q_\theta \mid \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^k$ for some $k \geq 1$, each q_θ defines a probability density or mass function on data $(X_1, \ldots, X_N) = X^N$ of sample size N, each outcome X_i taking a value in some space $\mathcal{X}$. The goal is to make inferences about θ, based on the data X^N or some statistic $S(X^N, N)$ thereof. In the common case with fixed $N = n$ and inference based on the full data, $S(X^N, N) := X^n$, we can transfer this statistical scenario to our setup by defining $\mathcal{P}^*$ as a set of distributions on $\mathcal{Z} = \Theta \times \mathcal{X}^n$. RVs U and $V = S(X^n, n) = X^n$ are then defined as, for each $z = (\theta, x^n)$, $U(z) := \theta$ and $V(z) := x^n$. DM employs a set Π of prior distributions on Θ, where each $\pi \in \Pi$ induces a joint distribution P_π on $\Theta \times \mathcal{X}^n$ with marginal on Θ determined by π and, given θ, density of x^n given by q_θ, so that if π has density p_π, we get the joint density $p_\pi(\theta, x^n) = p_\pi(\theta) \cdot q_\theta(x^n)$. We set $\mathcal{P}^* := \{P_\pi : \pi \in \Pi\}$ to be the set of all such joint distributions. In the special case in which DM really is a 100% subjective Bayesian who believes that a single prior π captures all uncertainty, we have that $\mathcal{P}^* = \{P_\pi\}$ contains just a single joint parameter-data distribution, and we are in the standard Bayesian scenario. Then DM can set $\tilde{P}(\theta \mid X^n) := P_\pi(\theta \mid X^n)$, the standard posterior, and any type of inference about θ is safe relative to $\mathcal{P}^*$. Here we focus on another special case, in which Π contains exactly one density for each $\theta \in \Theta$, namely the degenerate distribution putting all its mass on θ. We denote this distribution by P_θ and notice that then $\mathcal{P}^* = \{P_\theta : \theta \in \Theta\}$, with $P_\theta(\Theta = \theta) = 1$, and for any measurable set $\mathcal{A}$, $P_\theta(X^n \in \mathcal{A})$ determined by density p_θ, satisfying

$$p_\theta(x^n) = p_\theta(x^n \mid \Theta = \theta) = q_\theta(x^n).$$

Still, any choice of pragmatic distribution $\tilde{P}(U \mid V) = \tilde{P}(\theta \mid X^n)$ can be interpreted as a distribution on $U \equiv \Theta$ given the data X^n, analogous to a Bayesian posterior. In Section 3 we investigate how one can construct distributions $\tilde{P}$ of this kind that are safe for inference about *confidence intervals*. for simplicity we restrict ourselves to the 1-dimensional case, for which we find that the construction we provide leads to $\tilde{P}$ that are confidence-safe, written in our notation as 'safe for $\tilde{F}(U|V) \mid [V]$', with $\tilde{F}$ being the CDF (cumulative distribution function) of $\tilde{P}(U|V)$. Confidence safety is roughly the same as coverage Sweeting (2001): it means that the 'true' probability that θ is contained in a particular type of α-credible sets (sets with 'posterior' probability α given the data V), is equal to α.

The $\tilde{P}$ we construct are essentially equivalent to the *confidence distributions* of (Schweder and Hjort, 2002), that were designed with the explicit goal of having good confidence properties; they also often coincide with Fisher's 1930 fiducial distributions, which in later work (Fisher, 1935) he started treating as ordinary probability distributions that could be used without any restrictions. This cannot be right (see e.g. (Hampel, 2006, page 514)), but the question has always remained how a probability calculus for fiducial distributions could be derived that incorporates the right restrictions. Our work provides a step in this direction, in that we show how such $\tilde{P}$ snugly fit into our general framework: confidence safety is a strictly weaker property than calibration, and has again a natural representation in terms of the $\langle U \rangle \mid \langle V \rangle$ notation mentioned above. Moreover, it is a special case of *pivotal safety* which also has repercussions in quite different contexts — see Example 4.

The example illustrates two important points:

1. In some cases the literature suggests some method for constructing a pragmatic $\tilde{P}$. An example is the latent Dirichlet allocation model (Blei et al., 2003) mentioned above, in

which data V are text corpora, $\mathcal{P}^*$, not explicitly given, is a complicated set of realistic distributions over V under which data are non-i.i.d., and the literature suggests to take $\tilde{P}(U \mid V)$ as the Bayesian posterior for a cleverly designed i.i.d. model.

2. In other cases, DM may want to construct a $\tilde{P}$ herself. In Example 1, the safe $\tilde{P}$ was obtained by replacing an (unknown) conditional distribution with a (known) marginal — a special case of what was called $\mathcal{C}$-conditioning by Grünwald and Halpern (2011). Marginal distributions and distributions that ignore aspects of V play a more central role in this construction process: they also do in the confidence construction mentioned above, where one sets $\tilde{P}(U \mid V)$ equal to a distribution such that $\tilde{P}(U' \mid V)$, where U' is some auxiliary random variable (a *pivot*), becomes independent of V. For the original RV U though, in the dilation example, DM acts as if U and V are independent even though they may not be; in the confidence distribution example, DM acts in a 'dual' manner, namely as if U and V are dependent, even though under $\mathcal{P}^*$ they are not — which is fine, as long as her conclusions are *safe*.

Example 4 [Event-Based Conditioning and Pivotal Safety via Monty Hall] More generally, we may look at safety for pragmatic distributions $\tilde{P}$ that condition on events rather than random variables. To illustrate, consider the Monty Hall Problem (vos Savant, 1990, Gill, 2011): suppose that you're on a game show and given a choice of three doors $\{1, 2, 3\}$ Behind one is a car; behind the others are goats. You pick door 1. Before opening door 1, Monty Hall, the host opens one of the other two doors, say, door 3 which has a goat. He then asks you if you still want to take what's behind door 1, or to take what's behind door 2 instead. Should you switch? You may assume that initially, the car was equally likely to be behind each of the doors and that, after you go to door 1, Monty will always open a door with a goat behind. Basically you observe either the event $\{1, 3\}$ (if Monty opens door 2) or $\{1, 2\}$ (if Monty opens 3). You can then calculate your optimal decision according to some distribution $\tilde{P}(\{1\} \mid \mathcal{E})$, where $\mathcal{E} \in \{\{1, 3\}, \{1, 2\}\}$ is the event you observed. Naive conditioning suggests to take $P(\{1\} \mid \{1, 2\}) = P(\{1\} \mid \{1, 3\}) = (1/3)/(2/3) = 1/2$, and it takes a long time to convince most people that this is wrong — but, if DM's would adhere to safe probability, then no convincing and explanation would be needed: translation of the example into our 'safety' setting immediately shows, without any further thinking about the problem, that this choice of $\tilde{P}$ is *unsafe*, under all notions of safety we consider! (Section 4).

Another aspect of the Monty Hall problem is that, in most analyses that are usually viewed as 'correct', one implicitly assumes that the quiz master flips a *fair* coin to decide whether to open door 2 or 3 if you choose door 1 so that he has a choice. There have been heated discussions (e.g. on wikipedia talk pages) about whether this assumption is justified. In Example 11 we show that the $\tilde{P}$ which assumes a fair coin flip by Monty is an instance of a *pivotally safe* pragmatic distribution. These have the properties that for many loss functions (including 0/1-loss as in Monty Hall), they lead one to making optimal decisions. Thus, while assuming a fair coin flip may be wrong, it is still *harmless* to base one's decisions upon it.

Overview of the Paper In Section 2, we treat the case of countable space $\mathcal{Z}$, defining the basic notions of safety in Section 2.2 (where we return to dilation), and showing how calibration can be cleanly expressed using our notions in Section 2.3. In Section 3 we extend the setting to general $\mathcal{Z}$, which is needed to handle the case of confidence safety (Section 3.1), pivots (Section 3.2) and squared error optimality, where we observe continuous-valued random

variables. Section 4 briefly discusses non-numerical observations as well as probability updates that cannot be viewed as conditional distributions. We end with a discussion of further potential applications of safety as well as open problems. Proofs and further technical details are delegated to the appendix.

2 Basic Definitions for Discrete Random Variables

For simplicity, we introduce our basic notions only considering countable $\mathcal{Z}$, which allows us to sidestep measurability issues altogether. Thus below, $\mathcal{Z}$ is countable; we treat the general case in Section 3.

2.1 Concepts and notations regarding distributions on $\mathcal{Z}$

We define a random variable (abbreviated to RV) to be any function $X : \mathcal{Z} \to \mathbb{R}^k$ for some $k > 0$. Thus RVs can be multidimensional (i.e. what is usually called 'random vector'). By an '$\mathcal{Y}$-valued RV' or simply 'generalized RV' we mean any function mapping $\mathcal{Z}$ to an arbitrary set $\mathcal{Y}$. For two RVs $U = (U_1, \ldots, U_{k_1}), V = (V_1, \ldots, V_{k_2})$ where U_j and V_j are 1-dimensional random variables, we define (U, V) to be the RV with components $(U_1, \ldots, U_{k_1}, V_1, \ldots, V_{k_2})$.

For any generalized RVs U and V on $\mathcal{Z}$ and function f we write $U \overset{f}{\rightsquigarrow} V$ if for all $z \in \mathcal{Z}$, $V(z) = f(U(z))$. We write $U \rightsquigarrow V$ ("U determines V", or equivalently "U is a *coarsening* of V") if there is a function f such that $U \overset{f}{\rightsquigarrow} V$. We write $U \leftrightsquigarrow V$ if $U \rightsquigarrow V$ and $V \rightsquigarrow U$. For two GRVs U and V we write $U \equiv V$ if they define the same function on $\mathcal{Z}$, and for a distribution $P \in \mathcal{Z}$ we write $U =_P V$ if $P(U = V) = 1$. We write $U \overset{f}{\rightsquigarrow}_P V$ if $P(\{z \in \mathcal{Z} : V(z) = f(U(z))\}) = 1$, and $U \rightsquigarrow_P V$ if there exists some f for which this holds. Clearly $U \rightsquigarrow V$ implies that for all distributions P on $\mathcal{Z}$, $U \rightsquigarrow_P V$, but not vice versa. Let $S : \mathcal{Z} \to \mathcal{S}$ be a function on $\mathcal{Z}$. The *range* of S, denoted $\textsc{range}(S)$, the *support* of S under a distribution P, and the range of S given that another function T on $\mathcal{Z}$ takes value t, are denoted as

$$\begin{aligned} &\textsc{range}(S) := \{s \in \mathcal{S} : s = S(z) \text{ for some } z \in \mathcal{Z}\} \;\; ; \;\; \textsc{supp}_P(S) := \{s \in \mathcal{S} : P(S = s) > 0\}, \\ &\textsc{range}(S \mid T = t) = \{s \in \mathcal{S} : s = S(z) \text{ for some } z \in \mathcal{Z} \text{ with } t = T(z)\} \end{aligned} \tag{5}$$

where we note that $\textsc{supp}_P(S) \subseteq \textsc{range}(S)$, with equality if S has full support.

For a distribution P on $\mathcal{Z}$, and $\mathcal{U}$-valued RV U, we write $P(U)$ as short-hand to denote the distribution of U under P (i.e. $P(U)$ is a probability measure).

We generally omit double brackets, i.e. if we write $P(U, W)$ for RVs U and W, we really mean $P(R)$ where R is the RV (U, W),

Any generalized RV that maps all $z \in \mathcal{Z}$ to the same constant is called *trivial*, in particular the RV $\mathbf{0}$ which maps all $z \in \mathcal{Z}$ to 0. For an event $\mathcal{E} \subset \mathcal{Z}$, we define the *indicator random variable* $\mathbf{1}_{\mathcal{E}}$ to be 1 if $\mathcal{E}$ holds and 0 otherwise.

Conditional Distributions as Generalized RVs For given distribution on $\mathcal{Z}$ and generalized RVs V and W, we denote, for all $v \in \textsc{supp}_P(V)$, $P \mid V = v$ as the conditional distribution on $\mathcal{Z}$ given $V = v$, in the standard manner. We further define $(\mathcal{P}^* \mid W = w) := \{(P \mid W = w) : P \in \mathcal{P}^*, w \in \textsc{supp}_P(W)\}$ to be the set of distributions on $\mathcal{Z}$ that can be arrived at from $\mathcal{P}$ by conditioning on $W = w$, for all w supported by some $P \in \mathcal{Z}$.

We further denote, for all $v \in \text{SUPP}_P(V)$, $P(U \mid V = v)$ as the conditional distribution of U given $V = v$, defined as the distribution on U given by $P(U = u \mid V = v) := P(U = u, V = v)/P(V = v)$ (whereas $P \mid V = v$ is defined as a distribution on $\mathcal{Z}$, $P(U \mid V = v)$ is a distribution on the more restricted space $\text{RANGE}(U)$).

Suppose DM is interested in predicting RV U given RV V and does this using some conditional distribution $P(U \mid V = v)$ (usually this P will be the 'pragmatic' $\tilde{P}$, but the definition that follows holds generally). Adopting the standard convention for conditional expectation, we call any function from $\text{RANGE}(V)$ to the set of distributions on U that coincides with $P(U \mid V = v)$ for all $v \in \text{SUPP}_P(V)$ a *version* of the conditional distribution $P(U \mid V)$. If we make a statement of the form '$P(U \mid V)$ satisfies ...', we really mean 'every version of $P(U \mid V)$ satisfies...'. We thus treat $P(U \mid V)$ as a $\mathcal{E}$-valued random variable where $\mathcal{E} = \{P(U \mid V = v) : v \in \text{RANGE}(V)\}$, where, for all $z \in \mathcal{Z}$ with $P(V = V(z)) > 0$, $P(U \mid V)(z) := P(U \mid V = V(z))$, and $P(U \mid V)(z)$ set to an arbitrary value otherwise.

Unique and Well-Definedness Recall that DM starts with a set $\mathcal{P}^*$ of distributions on $\mathcal{Z}$ that she considers the right description of her uncertainty. She will predict sume RV U given some generalized RV V using a *pragmatic* distribution $\tilde{P}$.

For RV $U : \mathcal{Z} \to \mathbb{R}^k$ and generalized RV V, we say that, for given distribution P' on $\mathcal{Z}$, $P'(U \mid V)$ is *essentially uniquely defined* (relative to $\mathcal{P}^*$) if for all $P \in \mathcal{P}^*$, $\text{SUPP}_P(V) \subseteq \text{SUPP}_{P'}(V)$ (so that P-almost surely V takes value v with $P'(V = v) > 0$). We use this definition both for $P' \in \mathcal{P}^*$ and for $P' = \tilde{P}$; note that we always *evaluate* whether P' is uniquely defined under distributions in the 'true' $\mathcal{P}^*$ though.

We say that $\text{E}_{P'}[U \mid V]$ is well-defined if, writing $U = (U_1, \ldots, U_k)$, and, $U_j^+ = \max\{U_j, 0\}$, $U_j^- = \max\{-U_j, 0\}$, we have, for $j = 1..k$, either $\text{E}_{P'}[U_j^+ \mid V] < \infty$ with P-probability 1, or $\text{E}_{P'}[U_j^- \mid V] < \infty$ with P-probability 1. This is a very weak requirement that ensures that calculating expectations never involves the operation $\infty - \infty$, making all expectations well-defined.

The Pragmatic Distribution $\tilde{P}$ We assume that DM makes her predictions based on a probability distribution $\tilde{P}$ on $\mathcal{Z}$ which we generally refer to as the *pragmatic distribution*. In practice, DM will usually be presented with a decision problem in which she has to predict some fixed RV U based on some fixed RV V, and then she is only interested in the conditional distribution $\tilde{P}(U \mid V)$, and for some other RVs U' and V', $\tilde{P}(U' \mid V')$ may be left undefined. In other cases she only may want to predict the expectation of U given V — in that case she only needs to specify $\text{E}_{\tilde{P}}[U \mid V]$ as a function of V, and all other details of $\tilde{P}$ may be left unspecified. In Appendix A.1 we explain how to deal with such *partially specified* $\tilde{P}$. In the main text though, for simplicity we assume that $\tilde{P}$ is a fully-specified distribution on $\mathcal{Z}$; DM can fill up irrelevant details any way she likes. The very goal of our paper being to restrict $\tilde{P}$ to making 'safe' predictions however, DM may come up with $\tilde{P}$ to predict U given V and there may be many RVs U' and V' definable on the domain such that $\tilde{P}(U' \mid V')$ has no bearing to $\mathcal{P}^*$ and would lead to terrible predictions; as long as we make sure that $\tilde{P}$ is not used for such U' and V' — which we will — this will not harm the DM.

2.2 The Basic Notions of Safety

All our subsequent notions of 'safety' will be constructed in terms of the following first, simple definitions.

Definition 1 *Let $\mathcal{Z}$ be an outcome space and $\mathcal{P}^*$ be a set of distributions on $\mathcal{Z}$, let U be an RV and V be a generalized RV on $\mathcal{Z}$, and let $\tilde{P}$ be a distribution on $\mathcal{Z}$. We say that $\tilde{P}$ is* safe *for $\langle U \rangle \mid [\![V]\!]$ (pronounced as '$\tilde{P}$ is safe for* predicting $\langle U \rangle$ given $[\![V]\!]$*'), if*

$$\text{for all } P \in \mathcal{P}^* : \inf_{v \in \text{SUPP}_{\tilde{P}}(V)} \mathrm{E}_{\tilde{P}}[U|V=v] \leq \mathrm{E}_P[U] \leq \sup_{v \in \text{SUPP}_{\tilde{P}}(V)} \mathrm{E}_{\tilde{P}}[U|V=v]. \tag{6}$$

We say that $\tilde{P}$ is safe *for $\langle U \rangle \mid \langle V \rangle$, if*

$$\text{for all } P \in \mathcal{P}^* : \mathrm{E}_P[U] = \mathrm{E}_P[\mathrm{E}_{\tilde{P}}[U|V]]. \tag{7}$$

We say that $\tilde{P}$ is safe *for $\langle U \rangle \mid [V]$, if (6) holds with both inequalities replaced by an equality, i.e. for all $v \in \text{SUPP}_{\tilde{P}}(V)$,*

$$\text{for all } P \in \mathcal{P}^* : \mathrm{E}_P[U] = \mathrm{E}_{\tilde{P}}[U|V=v]. \tag{8}$$

In this definition, as in all definitions and results to come, whenever we write '$\langle$ statement $\rangle$' we really mean 'all conditional probabilities in the following statement are essentially uniquely defined, all expectations are well-defined, and $\langle$ statement $\rangle$'. Hence, (7) really means 'for all $P \in \mathcal{P}^*$, $\tilde{P}(U|V)$ is essentially uniquely defined, $\mathrm{E}_{\tilde{P}}[U|V]$, $\mathrm{E}_P[U]$, and $\mathrm{E}_P[\mathrm{E}_{\tilde{P}}[U|V]]$ are well-defined, and the latter two are equal to each other'. Also, when we wrote $\tilde{P}$ is safe for $\langle U \rangle \mid \langle V \rangle$, we really meant that it is safe for $\langle U \rangle \mid \langle V \rangle$ *relative to the given* $\mathcal{P}^*$; we will in general leave out the phrase 'relative to $\mathcal{P}^*$', whenever this cannot cause confusion.

To be fully clear about notation, note that in double expectations like in (7), we consider the right random variable to be bound by the outer expectation; thus it can be rewritten in any of the following ways:

$$\begin{aligned}
\mathrm{E}_{U \sim P}[U] &= \mathrm{E}_{V \sim P} \mathrm{E}_{U \sim \tilde{P}|V}[U] \\
\mathrm{E}_{V \sim P} \mathrm{E}_{U \sim P|V}[U] &= \mathrm{E}_{V \sim P} \mathrm{E}_{U \sim \tilde{P}|V}[U] \\
\sum_{u \in \text{RANGE}(U)} P(U=u) \cdot u &= \sum_{v \in \text{RANGE}(V)} P(V=v) \cdot \sum_{u \in \text{RANGE}(U)} \tilde{P}(U=u \mid V=v) \cdot u,
\end{aligned}$$

where the second equality follows from the tower property of conditional expectation.

Towards a Hierarchy It is immediately seen that, if $\tilde{P}$ is safe for $\langle U \rangle \mid [V]$, then it is also safe for $\langle U \rangle \mid \langle V \rangle$, and if it is safe for $\langle U \rangle \mid \langle V \rangle$, then it is also safe for $\langle U \rangle \mid [\![V]\!]$. Safety for $\langle U \rangle \mid [\![V]\!]$ is thus the weakest notion — it allows a DM to give valid upper- and lower-bounds on the actual expectation of U, by quoting $\sup_{v \in \text{SUPP}_{\tilde{P}}(V)} \mathrm{E}_{\tilde{P}}[U \mid V=v]$ and $\inf_{v \in \text{SUPP}_{\tilde{P}}(V)} \mathrm{E}_{\tilde{P}}[U \mid V=v]$, respectively, but nothing more. It will hardly be used here, except for a remark below Theorem 2; it plays an important role though in applications of safety to hypothesis testing, on which we will report in future work.

Safety for $\langle U \rangle \mid \langle V \rangle$ evidently bears relations to *unbiased* estimation: if $\tilde{P}$ is safe for $\langle U \rangle \mid \langle V \rangle$, i.e. (7) holds, then we can think of $\mathrm{E}_{\tilde{P}}[U|V]$ as an unbiased estimate, based on observing V, of the random quantity U (see also Example 8 later on). Safety for $\langle U \rangle \mid [V]$ implies that all distributions in $\mathcal{P}^*$ agree on the expectation of U and that $\mathrm{E}_{\tilde{P}}[U \mid V=v]$ is the same for (essentially) all values of v, and is thus a much stronger notion.

Example 5 [Dilation: Example 1, Cont.] The first application of definition (7) was already given in Example 1, where we used a $\tilde{P}$ that ignored V and was safe for $\langle U \rangle \mid \langle V \rangle$ and $\langle U \rangle \mid [V]$, as we see from (4) with g the identity. Let us extend the example, replacing $\mathcal{U} = \{0, 1\}$ in that example by $\mathcal{U} = \{0, 1, 2\}$, with $\mathcal{P}^*$ again defined as the set of all distributions satisfying (1) and $\tilde{P}$ defined by, for $v \in \{0, 1\}$, $\tilde{P}(U = 1 \mid V = v) = 0.9$, $\tilde{P}(U = 2 \mid V = v) = 0.09$. Then $\tilde{P}$ would still be safe for $\langle \mathbf{1}_{U=1} \rangle \mid \langle V \rangle$, but not for $\langle U \rangle \mid \langle V \rangle$: $\mathcal{P}^*$ contains a distribution whose marginal distribution $P(U = 2) = 0$, and (7) would not hold for that distribution.

Comparing the 'safety condition' (4) in Example 1 to (7) in Definition 1 we see that Definition 1 only imposes a requirement on expectations of U whereas (4) imposed a requirement also on RVs U' equal to functions $g(U)$ of U. For $\mathcal{U}$ with more than two elements as in Example 5 above, such a requirement is strictly stronger. We now proceed to define this stronger notion formally.

Definition 2 *Let $\mathcal{Z}, \mathcal{P}^*$, U, V and $\tilde{P}$ be as above. We say that $\tilde{P}$ is* safe *for $U \mid [\![V]\!]$ if for all RVs U' with $U \rightsquigarrow U'$, $\tilde{P}$ is safe for $\langle U' \rangle \mid [\![V]\!]$.*
Similarly, $\tilde{P}$ is safe for $U \mid \langle V \rangle$ if for all RVs U' with $U \rightsquigarrow U'$, $\tilde{P}$ is safe for $\langle U' \rangle \mid \langle V \rangle$, and $\tilde{P}$ is safe for $U \mid [V]$ if for all RVs U' with $U \rightsquigarrow U'$, $\tilde{P}$ is safe for $\langle U' \rangle \mid [V]$.

We see that safety of $\tilde{P}$ for $U \mid [V]$ implies that $\mathrm{E}_{\tilde{P}}[g(U) \mid V = v]$ is the same for all values of v in the support of $\tilde{P}$, and all functions g of U. This can only be the case if $\tilde{P}(U \mid V)$ *ignores* V, i.e. $\tilde{P}(U \mid V = v) = \tilde{P}(U)$, for all supported v. We must then also have that, for all $v \in \mathrm{SUPP}_{\tilde{P}}(V)$, that $\tilde{P}(U) = P(U)$, which means that all distributions in $\mathcal{P}^*$ agree on the marginal distribution of U, and $\tilde{P}(U)$ is equal to this marginal distribution. Thus, $\tilde{P}$ is safe for $U \mid [V]$ iff it is *marginally valid*. A prime example of such a $\tilde{P}(U \mid V)$ that ignores V and is marginally correct is the $\tilde{P}(U \mid V)$ we encountered in Example 1.

To get everything in place, we need a final definition.

Definition 3 *Let $\mathcal{Z}, \mathcal{P}^*$, U, V and $\tilde{P}$ be as above, and let W be another generalized RV.*

1. *We say that $\tilde{P}$ is* safe *for $\langle U \rangle \mid [\![V]\!], W$ if for all $w \in \mathrm{SUPP}_{\tilde{P}}(W)$, $\tilde{P} \mid W = w$ is safe for $\langle U \rangle \mid [\![V]\!]$ relative to $\mathcal{P}^* \mid W = w$. We say that $\tilde{P}$ is* safe *for $U \mid [\![V]\!], W$ if for all RVs U' with $U \rightsquigarrow U'$, $\tilde{P}$ is safe for $\langle U' \rangle \mid [\![V]\!], W$.*

2. *The same definitions apply with $[\![V]\!]$ replaced by $\langle V \rangle$ and $[V]$.*

3. *We say that $\tilde{P}$ is* safe *for $\langle U \rangle \mid W$ if it is safe for $\langle U \rangle \mid [\![\mathbf{0}]\!], W$; it is safe for $U \mid W$ if it is safe for $U \mid [\![\mathbf{0}]\!], W$.*

These definitions simply say that safety for '..|.., W' means that the space $\mathcal{Z}$ can be partitioned according to the value taken by W, and that for each element of the partition (indexed by w) one has 'local' safety given that one is in that element of the partition.

Proposition 1 gives reinterpretations of some of the notions above. The first one, (9) will mostly be useful for the proof of other results; the other three serve to make the original definitions more transparent:

Proposition 1 [Basic Interpretations of Safety] *Consider the setting above. We have:*

1. *$\tilde{P}$ is safe for $U \mid \langle V \rangle$ iff for all $P \in \mathcal{P}^*$, there exists a distribution P' on $\mathcal{Z}$ with for all $(u, v) \in \text{RANGE}((U, V))$, $P'(U = u, V = v) = \tilde{P}(U = u \mid V = v) \cdot P(V = v)$, that satisfies*

$$P'(U) = P(U). \tag{9}$$

2. *$\tilde{P}$ is safe for $\langle U \rangle \mid V$ iff for all $P \in \mathcal{P}^*$,*

$$\mathrm{E}_P[U \mid V] =_P \mathrm{E}_{\tilde{P}}[U \mid V]. \tag{10}$$

3. *$\tilde{P}$ is safe for $U \mid V$ iff for all $P \in \mathcal{P}^*$,*

$$P(U \mid V) =_P \tilde{P}(U \mid V). \tag{11}$$

4. *$\tilde{P}$ is safe for $U \mid [V], W$ iff for all $P \in \mathcal{P}^*$,*

$$P(U \mid W) =_P \tilde{P}(U \mid V, W). \tag{12}$$

Together with the preceding definitions, this proposition establishes the arrows in Figure 1 from $U \mid V$ to $\langle U \rangle \mid V$, from $\langle U \rangle \mid V$ to $\langle U \rangle \mid \langle V \rangle$ and from $U \mid \langle V \rangle$ to $\langle U \rangle \mid \langle V \rangle$. The remaining arrows will be established by Theorem 1 and 2.

Note that (12) says that $\tilde{P}$ is safe for $U \mid [V], W$ if $\tilde{P}$ ignores V *given* W, i.e. according to $\tilde{P}$, U is conditionally independent of V given W. Thus, $\tilde{P}$ can be safe for $U \mid [V], W$ and still $\tilde{P}(U \mid V)$ may depend on V; the definition only requires that V is ignored once W is given.

(11) effectively expresses that $\tilde{P}(U \mid V)$ is valid (a frequentist might say 'true') for predicting U based on observing V, where as always we assume that $\mathcal{P}^*$ itself correctly describes our beliefs or potential truths (in particular, if $\mathcal{P}^* = \{P\}$ is a singleton, then any $\tilde{P}(U \mid V)$ which coincides a.s. with $P(U \mid V)$ is automatically valid). Thus, 'validity for $U \mid V$', to be interpreted as *$\tilde{P}$ is a valid distribution to use when predicting U given observations of V* is a natural name for safety for $U \mid V$. We also have a natural name for safety for $\langle U \rangle \mid V$: for 1-dimensional U, (10) simply expresses that all distributions in $\mathcal{P}^*$ agree on the conditional expectation of $U \mid V$, and that $\mathrm{E}_{\tilde{P}}[U \mid V]$ is a version of it. which implies (see e.g. Williams (1991)) that, with the function $g(v) := \mathrm{E}_{\tilde{P}}[U \mid V = v]$,

$$\mathrm{E}_{(U,V)\sim P}[(U - g(V))^2] = \min_f \mathrm{E}_{(U,V)\sim P}[(U - f(V))^2], \tag{13}$$

the minimum being taken over all functions from $\text{RANGE}(V)$ to $\mathbb{R}$. This means that $\tilde{P}$ encodes the *optimal* regression function for U given V and hence suggests the name *squared-error optimality*. Summarizing the names we encountered (see Figure 1):

Definition 4 [(Potential) Validity, Squared Error-Optimality, Unbiasedness, Marginal Validity] *If $\tilde{P}$ is safe for $U \mid V$, i.e. (11) holds for all $P \in \mathcal{P}^*$, then we also call $\tilde{P}$* valid *for $U \mid V$ (again, pronounce as 'valid for* predicting *U* given *V'). If (11) holds for* some *$P \in \mathcal{P}^*$, we call $\tilde{P}$* potentially valid *for $U \mid V$. If $\tilde{P}$ is safe for $\langle U \rangle \mid V$, we call $\tilde{P}$* squared error-optimal *for $U \mid V$. If $\tilde{P}$ is safe for $\langle U \rangle \mid \langle V \rangle$, we call $\tilde{P}$* unbiased *for $U \mid V$. If $\tilde{P}$ is safe for $\langle U \rangle \mid [V]$, we say that it is* marginally valid *for $U \mid V$.*

It turns out that there also is a natural name for safety for $U \mid [V], W$ whenever $V \rightsquigarrow W$. The next example reiterates its importance, and the next section will provide the name: *calibration*.

Example 6 Suppose $\tilde{P}$ is safe for $U \mid [V_1], V_2$. From Proposition 1, (12) we see that this means that for all $P \in \mathcal{P}^*$, all $v_1, v_2 \in \text{SUPP}_P(V_1, V_2)$, that

$$\text{E}_P[U' \mid V_2 = v_2] = \text{E}_{\tilde{P}}[U' \mid V_1 = v_1, V_2 = v_2], \tag{14}$$

The special case with $V_2 \equiv \mathbf{0}$ has already been encountered in Example 1, (3). As discussed in that example, for $V_2 \equiv \mathbf{0}$, (14) expresses our basic interpretation of safety that *predictions based on $\tilde{P}$ will always be as good, in expectation, as the DM who uses $\tilde{P}$ expects them to be.* Clearly this continues to be the case if (14) holds for some nontrivial V_2.

2.3 Calibration Safety

In this section, we show that *calibration*, as informally defined in Example 2, has a natural formulation in terms of our safety notions. We first define calibration formally, and then, in our first main result, Theorem 1, show how being calibrated for predicting U based on observing V is essentially equivalent to being safe for $U \mid [V], V'$ for some types of V' that need not be equal to V itself, including $V' \equiv \mathbf{0}$. Thus, we now effectively unify the ideas underlying Example 1 (dilation) and Example 2 (calibration).

Following Grünwald and Halpern (2011) we define calibration directly in terms of distributions rather than empirical data, in the following way:

Definition 5 [Calibration] *Let $\mathcal{Z}$, $\mathcal{P}^*$, U, V and $\tilde{P}$ be as above. We say that $\tilde{P}$ is* calibrated *(or* calibration–safe*) for $\langle U \rangle \mid V$ if for all $P \in \mathcal{P}^*$, all $\mu \in \{\text{E}_{\tilde{P}}[U \mid V = v] : v \in \text{SUPP}_P(V)\}$,*

$$\text{E}_P[U \mid \ \text{E}_{\tilde{P}}[U \mid V] = \mu \ ] = \mu. \tag{15}$$

We say that $\tilde{P}$ is calibrated *for $U \mid V$ if for all $P \in \mathcal{P}^*$, all $p \in \{\tilde{P}(U \mid V = v) : v \in \text{SUPP}_P(V)\}$,*

$$P(U \mid \ \tilde{P}(U \mid V) = p \ ) = p \tag{16}$$

Hence, calibration (for U) means that given that a DM who uses $\tilde{P}$ *predicts* a specific distribution for U, the *actual* distribution is indeed equal to the predicted distribution. Note that here we once again treat $\tilde{P}(U \mid V)$ as a generalized RV.

In practice we would want to weaken Definition 5 to allow some slack, requiring the μ (viz. p) inside the conditioning to be only within some $\epsilon > 0$ of the μ (viz. p) outside, but the present idealized definition is sufficient for our purposes here. Note also that the definition refers to a simple form of calibration, which does not involve selection rules based on past data such as used by, e.g., Dawid (1982).

We now express calibration in terms of our safety notions. We will only do this for the 'full distribution'–version (16); a similar result can be established for the average-version.

Theorem 1 *Let U, V and $\tilde{P}$ be as above. The following three statements are equivalent:*

1. *$\tilde{P}$ is calibrated for $U \mid V$;*
2. *There exists a RV V' on $\mathcal{Z}$ with $V \rightsquigarrow_{\tilde{P}} V'$ such that $\tilde{P}$ is safe for $U \mid [V], V'$*

3. *$\tilde{P}$ is safe for $U \mid V''$ where V'' is the generalized RV given by $V'' \equiv \tilde{P}(U \mid V)$.*

Note that, since safety for $U \mid V$ implies safety for $U \mid [V], V'$ for $V' = V$, *(2.)* $\Rightarrow$ *(1.)* shows that safety for $U \mid V$ implies calibration for $U \mid V$. By mere definition chasing (details omitted) one also finds that *(2.)* implies that $\tilde{P}$ is safe for $U \mid \langle V \rangle, V'$ and, again by definition chasing, that $\tilde{P}$ is safe for $U \mid \langle V \rangle$. Thus, this result establishes two more arrows of the hierarchy of Figure 1. Its proof is based on the following simple result, interesting in its own right:

Proposition 2 *Let V and V' be generalized RVs such that $V \overset{f}{\rightsquigarrow}_{\tilde{P}} V'$ for some function f. The following statements are equivalent:*

1. *$\tilde{P}(U \mid V, V')$ ignores V, i.e. $\tilde{P}(U \mid V, V') =_{\tilde{P}} \tilde{P}(U \mid V')$.*
2. *For all $v' \in \text{SUPP}_{\tilde{P}}(V')$, for all $v \in \text{SUPP}_{\tilde{P}}(V)$ with $f(v) = v'$: $\tilde{P}(U \mid V = v) = P(U \mid V' = v')$.*
3. *$V' \rightsquigarrow_{\tilde{P}} \tilde{P}(U \mid V)$*
4. *$V' \rightsquigarrow_{\tilde{P}} V''$ and $\tilde{P}(U \mid V', V'')$ ignores V', where $V'' = \tilde{P}(U \mid V)$.*

Moreover, if $\tilde{P}$ is safe for $U \mid V$ and $\tilde{P}(U \mid V, V')$ ignores V, then $\tilde{P}$ is safe for $U \mid V'$.

3 Continuous-Valued U and V; Confidence and Pivotal Safety

Our definitions of safety were given for countable $\mathcal{Z}$, making all random variables involved have countable range as well. Now we allow general $\mathcal{Z}$ and hence continuous-valued U and general uncountable V as well, but we consider a version of safety in which we do not have safety for $U \mid V$ itself, but for $U' \mid V$ for some U' with $(U, V) \rightsquigarrow U'$ such that the range of $\tilde{\mathcal{P}}_{[V]}(U') := \{\tilde{P}(U' \mid V = v) : v \in \text{RANGE}(V)\}$ is still countable. To make this work we have to equip $\mathcal{Z}$ with an appropriate σ-algebra $\Sigma_{\mathcal{Z}}$ and have to add to the definition of a RV that it must be measurable,[6] and we have to modify the definition of support $\text{SUPP}_P(U)$ to the standard measure-theoretic definition (which specializes to our definition (5) whenever there exists a countable $\mathcal{U}$ such that $P(U \in \mathcal{U}) = 1$). Yet nothing else changes and all previous definitions and propositions can still be used.[7]

Additional Notations and Assumptions In this section we frequently refer to (cumulative) distribution functions of 1-dimensional RVs, for which we introduce some notation: for distribution $P \in \mathcal{P}^*$ and RV $W : \mathcal{Z} \to \mathbb{R}$, let $F_{[W]}(\cdot)$ denote the distribution function of W, i.e. $F_{[W]}(w) = P(W \leq w)$. The notation is extended to conditional distribution functions: for given $\tilde{P}(U \mid V)$, we let $\tilde{F}_{[U|V]}(u|v) := \tilde{P}(U \leq u \mid V = v)$. The subscripts $[W]$ and $[U|V]$ indicate the RVs under consideration; we will omit them if they are clear from

[6] Formally we assume that $\mathcal{Z}$ is equipped with some σ-algebra $\Sigma_{\mathcal{Z}}$ that contains all singleton subsets of $\mathcal{Z}$. We associate the co-domain of any function $X : \mathcal{Z} \to \mathbb{R}^k$ with the standard Borel σ-algebra on $\mathbb{R}^k$, and we call such X an RV whenever the σ-algebra $\Sigma_{\mathcal{Z}}$ on $\mathcal{Z}$ is such that the function is measurable.

[7] If we were to consider safety of the form $U \mid V$ for uncountable $\tilde{\mathcal{P}}_{[V]}(U)$, then this set of probability distributions would have to be equipped with a topology, which is a bit more complicated and is left for future work.

the context. Note that we can consider these distribution functions as RVs: for all $z \in \mathcal{Z}$, $F_{[W]}(W)(z) = P(W \leq W(z))$ and $\tilde{F}_{[U|V]}(U|V)(z) = \tilde{P}(U \leq U(z) \mid V = V(z))$.

Since this greatly simplifies matters, we will often assume that either $P \in \mathcal{P}^*$ or $\tilde{P}(U \mid V)$ satisfy the following:

Scalar Density Assumption A distribution $P(U)$ for RV U satisfies the *scalar density assumption* if (a) $\text{RANGE}(U) \subseteq \mathbb{R}$ is equal to some (possibly unbounded) interval, and (b) P has a density f relative to Lebesgue measure with $f(u) > 0$ for all u in the interior of $\text{RANGE}(U)$. We say that $P(U|V)$ satisfies the scalar density assumption if for all $v \in \text{RANGE}(V)$, $P(U \mid V = v)$ satisfies it.

This is a strong assumption which will nevertheless be satisfied in many practical cases. For example, normal distributions, gamma distributions with fixed shape parameter, beta distributions etc. all satisfy it.

Overview of this Section The goal of the following two subsections is to precisely reformulate the *fiducial* and *confidence* distributions that have been proposed in the statistical literature as pragmatic distributions in our sense, that can be safely used for some ('confidence-related') but not for other prediction tasks. Here we focus on the standard statistical scenario introduced in Example 3. The underlying idea of 'pivotal safety' (developed in Section 3.2) has applications in discrete, nonstatistical settings as well, as explored in Section 3.3.

3.1 Confidence Safety

We start with a classic motivating example.

Example 7 [Example 3, Specialized] As a special case of the statistical scenario outlined in Example 3, let $\mathcal{M}$ be the normal location family with varying mean θ and fixed variance, say $\sigma^2 = 1$, and let $V := \hat{\theta} = \hat{\theta}(X^n)$ where $\hat{\theta}(X^n)$ is the empirical average of the X_i, which is a sufficient statistic that is of course also equal to the ML estimator for data X^n. Then the sampling density of $\hat{\theta}$ is itself Gaussian, and given by

$$p(\hat{\theta} \mid \theta) \propto q_\theta(X^n) \propto e^{-\frac{1}{2} \cdot n \cdot (\hat{\theta} - \theta)^2}. \tag{17}$$

In this simple context, Fisher's controversial fiducial reasoning amounts to observing that (17) is symmetric in $\hat{\theta}$ and θ; thus, if we simply define a new function $\tilde{p}(\theta \mid \hat{\theta}) := p(\hat{\theta} \mid \theta)$, then this function must, for each fixed $\hat{\theta}$, be the density of a probability distribution (the integral over θ must by symmetry be 1); and this would then amount to something like a 'prior–free' posterior for θ based on data $\hat{\theta}$. In this special case, as well as with the corresponding inversion for scale families, it coincides with the Bayes posterior based on an improper Jeffreys' prior. Yet, Lindley (1958) showed that the general construction for 1-dimensional families, which we review in the next subsection, *cannot* correspond to a Bayesian posterior except for location and scale families: for different sample sizes, the 'fiducial' posterior for data of size n corresponds to the Bayes posterior for a prior which depends on n.

Fisher (1930) noted that $\tilde{p}$ as constructed above lead to valid inference about confidence intervals. Later (Fisher, 1935) he made claims that $\tilde{p}$ could be used for general prior-free

inference about θ given data/statistic $\hat{\theta}$. This is not correct though, and more recently, $\tilde{p}$ is more often regarded as an instance of a *confidence distribution* (Schweder and Hjort, 2002), a term going back to Cox (1958) — these are by and large the same objects as fiducial distributions, though with a stipulation that they only be used for certain inferences related to confidence. In the remainder of this subsection, we develop a variation of safety that can capture such confidence statements. In the next subsection, we review the general method for designing confidence distributions for 1-dimensional statistical families and we shall see that, under an additional condition, they are indeed confidence–safe in our sense. In the remainder of this section, we focus on 1-dimensional families and interpret the RVs U and V as in our statistical application of Example 3 and 7. Thus, $U \equiv \theta$ would be a 1-dimensional scalar parameter of some model $\{P_\theta : \theta \in \Theta\}$, $V \equiv S(X^n)$ would be a statistic of the observed data. In Example 7, $V \equiv \hat{\theta}(X^n)$ is the ML estimator.

We are thus interested in constructing, for each $v \in \text{RANGE}(V)$, an interval of $\mathbb{R}$ that has (say) 95% probability under $\tilde{P}(U \mid V = v)$. To this end, we define for each $v \in \text{RANGE}(V)$, an interval $\mathcal{C}_v = [\underline{u}_v, \bar{u}_v]$ where $\underline{u}_v$ is such that $\tilde{F}_{[U|V]}(\underline{u}_v \mid v) = 0.025$ and $\bar{u}_v$ is such that $\tilde{F}_{[U|V]}(\bar{u}_v \mid v) = 0.975$. This set obviously has 95% probability according to $\tilde{P}(U \mid V = v)$. In our interpretation where $U = \theta$ is the parameter of a statistical model, we may interpret $\tilde{P}$ as DM's assessment, given data $V = S(X^n)$, of the uncertainty about U, i.e. $\tilde{P}$ is a 'posterior' and, analogous to Bayesian terminology, we may call $\mathcal{C}_v$ a 95% *credible set* given V. The question is now under what conditions we have *coverage*, i.e. that $\mathcal{C}_V$ is also a 95% frequentist *confidence interval*, so that our credible set can be given frequentist meaning. By definition of confidence interval, this will be the case iff for all $P \in \mathcal{P}^*$, $P(U \in \mathcal{C}_V) = 0.95$, i.e. iff for all $P \in \mathcal{P}^*$, $v \in \text{RANGE}(V)$,

$$\mathrm{E}_P[\mathbf{1}_{U \in \mathcal{C}_V}] = \mathrm{E}_{\tilde{P}}[\mathbf{1}_{U \in \mathcal{C}_V} \mid V = v], \tag{18}$$

where we used that, by construction, $\mathrm{E}_{\tilde{P}}[\mathbf{1}_{U \in \mathcal{C}_V} \mid V = v] = 0.95$ for all $v \in \text{RANGE}(V)$. As we shall see (18) holds for our normal example, so the posterior constructed in (17) produces valid confidence intervals. (18) is of the form of a 'safety' statement and it suggests that confidence interval validity of credible sets can be phrased in terms of safety in general. Indeed this is possible as long as $\tilde{P}(U|V)$ satisfies the scalar density assumption: for fixed $0 \leq a < b \leq 1$, we can define the set $\mathcal{C}_v^{[a,b]} = [\underline{u}_v^a, \bar{u}_v^b]$ where $\tilde{F}_{[U|V]}(\underline{u}_v^a \mid v) = a$ and $\tilde{F}_{[U|V]}(\bar{u}_v^b \mid v) = b$, so that for each $v \in \text{RANGE}(V)$, $\mathcal{C}_v^{[a,b]}$ is a $b - a$ credible set. Reasoning like above, we then get that $\mathcal{C}_V^{[a,b]}$ is also a $b - a$ confidence interval iff for all $P \in \mathcal{P}^*$, all $v \in \text{RANGE}(V)$

$$\mathrm{E}_{(U,V)\sim P}[\mathbf{1}_{U \in \mathcal{C}_V^{[a,b]}}] = \mathrm{E}_P \mathrm{E}_{\tilde{P}}[\mathbf{1}_{U \in \mathcal{C}_V^{[a,b]}} \mid V = v], \tag{19}$$

which, from the characterization of safety for $U \mid [V]$, Proposition 1, (12) and (14) suggests the following definition:

Definition 6 *Let U, V and $\tilde{P}$ be such that $\tilde{P}(U|V = v)$ satisfies the scalar density assumption for all $v \in \text{RANGE}(V)$. We say that $\tilde{P}$ is (strongly)* confidence–safe *for $U \mid V$ if for all $0 \leq a < b \leq 1$, it is safe for $\mathbf{1}_{U \in \mathcal{C}_V^{[a,b]}} \mid [V]$.*

The requirement that $\tilde{P}$ satisfies the scalar density assumption is imposed because otherwise $\mathcal{C}_V^{[a,b]}$ may not be defined for some $a, b \in [0, 1]$. We could also consider distributions that have coverage in a slightly weaker sense, and define weak confidence-safety for $U \mid V$ as safety

for $\mathbf{1}_{U \in \mathcal{C}_V^{[a,b]}} \mid \langle V \rangle$; we have not (yet) found any natural examples though that exhibit weak confidence-safety but not strong confidence safety.

Example 8 In the next subsection we show that $\tilde{P}(\theta \mid V = \hat{\theta}(X^n))$ as defined in Example 7 (normal distributions) is confidence-safe. For example, we may specify a $\tilde{P}(\theta \mid \hat{\theta})$–95% credible set $\mathcal{C}_V^{[a,b]}$ with $a = 0.025$ and $b = 0.975$ as the area under the normal curve centered at $V = \hat{\theta}$ and truncated so that the area under the left and right remaining tails is 0.025 each. Now suppose that $X^n \sim P_\theta$ for arbitrary θ. By confidence–safety we know that the probability that we will observe $\hat{\theta}$ such that $\theta \notin \mathcal{C}_V^{[a,b]}$ is exactly 0.05, just as it would be if $\tilde{P}$ where the true conditional distribution — an instance of a *safe* inference based on $\tilde{P}$. For an example of an inference that is *unsafe*, suppose DM really is offered a gamble for \$1 that pays out \$2 whenever $\theta > 0$ (we could take any other fixed value as well), and pays out 0 otherwise. She thus has two actions at her disposal, $a = 1$ (accept the gamble) and $a = 0$ (abstain), with loss given by $L(\theta, 0) = 0$ for all θ and $L(\theta, 1) = 1$ if $\theta < 0$ and $L(\theta, 1) = -1$ otherwise. She might thus be tempted to follow the decision rule $\delta(\hat{\theta})$ that accepts the gamble whenever she observes $\hat{\theta}$ such that $\tilde{P}(\theta > 0 \mid \hat{\theta}) > .5$ and abstains otherwise; for that rule minimizes, among all decision rules, her expected loss $\mathrm{E}_{\theta \sim \tilde{P}|\hat{\theta}}[L(\theta, \delta(\hat{\theta}))]$, and gives negative expected loss.

This decision rule should not be followed though, because it is based on an inference that is not safe in any of our senses: safety would mean that $\tilde{P}$ is safe for $L(\theta, \delta(\hat{\theta})) \mid \mathrm{s}$, where s can be substituted by $[\hat{\theta}]$, $\langle \hat{\theta} \rangle$, or $\hat{\theta}$. The first does not apply since $\hat{\theta}$ is not ignored in the probability assessment; the third does not hold because it would imply the second, which also does not hold. To see this, note that if data comes from $P_{\bar{\theta}}$ with $\bar{\theta} < 0$ then we have

$$\mathrm{E}_{\hat{\theta} \sim P_{\bar{\theta}}}[L(\bar{\theta}, \delta(\hat{\theta}))] > 0 > \mathrm{E}_{\hat{\theta} \sim P_{\bar{\theta}}}[\mathrm{E}_{\theta \sim \tilde{P}|\hat{\theta}}[L(\theta, \delta(\hat{\theta}))]],$$

so that her actual expected loss is positive whereas she thinks it to be negative. This violates (7) in Definition 1 so that $\tilde{P}$ is not safe for $L(\theta, \delta(\hat{\theta})) \mid \langle \hat{\theta} \rangle$. Note that, if $\tilde{P}$ were safe for $\theta \mid \hat{\theta}$ (as a subjective Bayesian would believe if $\tilde{P}$ were her posterior) then it would also be safe for $L(\theta, \delta(\hat{\theta}))$ (because $L(\theta, \delta(\hat{\theta}))$ can be written as a function of $(\theta, \hat{\theta})$), and then use of δ would be safe after all.

For an intuitive interpretation, consider a long sequence of experiments. For each j, in the j-th experiment, a sample of size $n = 10$ is drawn from a normal with some mean θ_j. Each time DM investigates whether $\theta_j > 0$. Assume that, in reality, all of the θ_j are < 0, but DM does not know this. Then every once in a while $\hat{\theta}$ will be large enough for our unsafe DM to gamble on it, but every time this happens she loses; all other times she neither loses nor wins, so her net gain is negative in the long run.

Thus, $\tilde{P}$ is not safe for $\theta \mid \hat{\theta}$ in general. However, it is still safe for $U'|V$ for some other functions of $U \equiv \theta$ besides $U' = \mathcal{C}_V^{[a,b]}$. For example, it leads to unbiased estimation of the mean: $\tilde{P}$ is safe for $\langle \theta \rangle \mid \langle \hat{\theta} \rangle$, as is easily established. This is however a special property of the confidence distribution for the normal location family and does not hold for general 1-dimensional confidence distributions as reviewed below.

3.2 Pivotal Safety and Confidence

Trivially, if $\tilde{P}$ is safe for $U|V$ (hence valid) and the scalar density assumption holds, then it is also confidence-safe for $U|V$. We now determine a way to construct confidence-safe $\tilde{P}$

if not enough knowledge is available to infer a $\tilde{P}$ that is valid. To this end, we invoke the concept of a *pivot*, usually defined as a function of the data and the parameter that has the same distribution for every $P \in \mathcal{P}^*$ and that is monotonic in the parameter for every possible value of the data (Barndorff-Nielsen and Cox, 1994). We adopt the following variation that also covers a quite different situation with discrete outcomes:

Definition 7 **[pivot]** *Let U and V be as before and suppose either (continuous case) that $U : \mathcal{Z} \to \mathbb{R}$ and $V : \mathcal{Z} \to \mathbb{R}$ are real-valued RVs, and that for all $v \in \text{RANGE}(V)$, $\text{RANGE}(U|V = v)$ is a (possibly unbounded) interval (possibly dependent on v), or (discrete case) that $\mathcal{Z}$ is countable. We call RV U' a (continuous viz. discrete)* pivot *for $U \mid V$ if*

1. *$(U, V) \rightsquigarrow U'$ so that the function f with $U' = f(U, V)$ exists.*
2. *For each fixed $v \in \text{RANGE}(V)$, the function $f_v : \text{RANGE}(U|V = v) \to \text{RANGE}(U')$, defined as $f_v(u) := f(u, v)$ is 1-to-1 (an injection); in the continuous case we further require f_v to be continuous and uniformly monotonic, i.e. either $f_v(u)$ is increasing in u for all $v \in \text{RANGE}(V)$, or $f_v(u)$ is decreasing in u for all $v \in \text{RANGE}(V)$.*
3. *All $P \in \mathcal{P}^*$ agree on U', i.e. for all $P_1, P_2 \in \mathcal{P}^*$, $P_1(U') = P_2(U')$, where in the continuous case we further require that P_1 (hence also P_2) satisfies the scalar density assumption.*

We say that a pivot U' is simple *if for all $v \in \text{RANGE}(V)$, the function f_v is a bijection.*

The scalar density assumption (item 3) does not belong to the standard definition of pivot, but it is often assumed implicitly, e.g. by Schweder and Hjort (2002). The importance of 'simple' pivots (a nonstandard notion) will become clear below.

In the remainder of this section we focus on the statistical case of the previous subsection, which is a special case of Definition 7 above — thus invariably $U \equiv \theta$, the 1-dimensional parameter of a model $\{P_\theta \mid \theta \in \Theta\}$, and V is some statistic of data X^n. In Section 3.3 we return to the discrete case.

If a continuous pivot as above exists, then all $P \in \mathcal{P}^*$ have the same distribution function $F_{[U']}(u') := P(U' \leq u')$. We may thus define a pragmatic distribution by setting, for all $v \in \text{RANGE}(V)$, all $u \in \text{RANGE}(U \mid V = v)$,

$$\tilde{F}_{[U|V]}(u \mid v) := \begin{cases} F_{[U']}(f_v(u)) & \text{if } f_v(u) \text{ increasing in } u \\ 1 - F_{[U']}(f_v(u)) & \text{if } f_v(u) \text{ decreasing.} \end{cases} \tag{20}$$

The definition of pivot ensures that for each $v \in \text{RANGE}(V)$, $\tilde{F}_{[U|V]}(u \mid v)$ is a continuous increasing function of u that is in between 0 and 1 on all $u \in \text{RANGE}(U \mid V = v)$, and hence $\tilde{F}_{[U|V]}(u \mid v)$ is the CDF of some distribution $\tilde{P}(U|V)$. It can be seen from the standard definition of a confidence distribution (Schweder and Hjort, 2002) that this $\tilde{P}(U|V)$ is a confidence distribution, and that every confidence distribution can be obtained in this way.[8]

[8] Mirroring the discussion underneath Definition 1 from Schweder and Hjort (2002): if $\tilde{F}'(U|V)$ is the CDF of a confidence distribution as defined by them, then $U' := \tilde{F}'(U|V)$ is a pivot and then the construction above applied to U' gives $\tilde{F}(U|V) := \tilde{F}'(U|V)$. Conversely, if U' is an arbitrary continuous pivot, then by the requirement that $P(U')$ has a density with interval support, $F(U')$ is itself uniformly distributed on its support $[0, 1]$ and there is a 1-to-1 continuous mapping between U' and $F(U')$. Thus, whenever U' is a continuous pivot, $F(U')$ is itself a pivot as well, and $\tilde{F}(U|V)$ as defined here satisfies the definition of confidence distribution.

Hence, (20) essentially defines confidence distribution. Theorem 2 below shows that when based on *simple* pivots, confidence distributions are also confidence-safe.

Example 9 Consider the statistical setting with $U \equiv \theta$, $V \equiv \hat{\theta}(X^n)$, and (a) for all $\theta \in \Theta$, $P_\theta(V)$ itself satisfies the scalar density assumption, and (b) for each fixed $v \in \text{RANGE}(V)$, we have that $F_{\theta,[V]}(v) := P_\theta(\hat{\theta}(X^n) \leq v)$ is monotonically decreasing in θ. This will hold for 1-dimensional exponential families with a continuously supported sufficient statistic (such as the normal, exponential, beta- and many other models), taken in their mean-value parameterization Θ. Then (by (b)) $U' = F_{\theta,[V]}(V)$ is itself a decreasing pivot, with (by (a)) the additional property that the function f_θ from $\text{RANGE}(V)$ to $\text{RANGE}(U')$ given by $f_\theta(v) := f(\theta, v)$ is strictly increasing in v. Then (20) simplifies, because (using this strict increasingness in the second equality):

$$F_{\theta,[V]}(v) = P_\theta(V \leq v) = P_\theta(F_{\theta,[V]}(V) \leq f_\theta(v)) = F_{\theta,[F_{\theta,[V]}]}(f_\theta(v)) = F_{[U']}(f_v(\theta)),$$

and noticing that the right-hand side appears in (20), we can plug in the left-hand side there as well and we see that we can directly set

$$\tilde{F}(\theta \mid \hat{\theta}) = 1 - F_\theta(\hat{\theta}). \tag{21}$$

Thus for such models the recipe (20) simplifies (see also Veronese and Melilli (2015)).

We now define 'pivotal safety' which, as demonstrated below, in the statistical case essentially coincides with confidence safety — the reason for the added generality is that it also has meaning and repercussions in the discrete case. The extension to 'multipivots' is just a stratification that means that, given any $w \in \text{RANGE}(W)$, $\tilde{P} \mid W = w$ is pivotally safe for $U \mid V$ relative to $\mathcal{P}^* \mid W = w$; it is not really needed in this text, but is convenient for completing the hierarchy in Figure 1.

Definition 8 *Let U and V be as before and let $\tilde{P}$ be an arbitrary distribution on $\mathcal{Z}$ (not necessarily given by (20)) . If V has full support under $\tilde{P}$, i.e. $\text{SUPP}_{\tilde{P}}(V) = \text{RANGE}(V)$ and U' is a (continuous or discrete) pivot such that $\tilde{P}$ is safe for $U' \mid [V]$, i.e. for all $v \in \text{RANGE}(V)$,*

$$\tilde{P}(U' \mid V = v) = \tilde{P}(U'),$$

then we say that $\tilde{P}$ is pivotally safe for $U \mid V$, with pivot U'.

Now let W be a generalized RV such that $V \rightsquigarrow W$. Suppose that for all $w \in \text{RANGE}(W)$, U' is a pivot relative to the set of distributions $(\mathcal{P}^* \mid W = w)$ and $\tilde{P}$ is safe for $U' \mid [V], W$. Then we say that $\tilde{P}$ is pivotally safe for $U \mid V$ with multipivot *$U|W$.*

Example 10 [normal distributions and general confidence distributions] In Example 7, $U = \theta$, the mean of a normal with variance 1, and we set $V = \hat{\theta}(X^n)$ to be the average of a sample of size n. Then it is easily seen that $U' = \theta - \hat{\theta} = U - V$ is a pivot according to our definition, having a $N(0,1)$ distribution under all $P \in \mathcal{P}^*$. If we adopt $\tilde{P}(U \mid V)$, under which $U' \sim N(0,1)$ independently of V, then $\tilde{P}$ is safe for $U' \mid [V]$ (see Proposition 1, (12)) so we have pivotal safety. And indeed one can verify that this $\tilde{P}$ coincides with the recipe given by (20).

While in the simplest form of calibration, safety for $U \mid [V]$, we had that $\tilde{P}(U \mid V)$ was the marginal of U, so that U and V are independent under $\tilde{P}$, in the simplest pivotal safety case, the situation is comparable, but now the auxiliary variable U' instead of the original variable U is independent of V under $\tilde{P}$. Note though that we do not necessarily have that U' and V are independent for all or even for *any* $P \in \mathcal{P}^*$. In Example 11 (Monty Hall) below, there is in fact just one single $P \in \mathcal{P}^*$ for which $U' \perp V$ holds. In the statistical application, we even have that $P(U' \mid V)$ is a degenerate distribution (putting all its mass on a particular real number depending on V) under each $P \in \mathcal{P}^*$, as can be checked from Example 7.

The relation between pivotal, confidence-safety and the $\tilde{P}$ defined as in (20) is given by the following central result.

Theorem 2 *The following statements are all equivalent:*

1. *$\tilde{P}$ is pivotally safe for $U|V$ with some* simple *continuous pivot U'*
2. *$\tilde{P}(U|V)$ is of form (20), where U' is a simple continuous pivot*
3. *$\tilde{P}$ is confidence-safe for $U|V$ and for each $v \in \text{RANGE}(V)$, $\tilde{P}(U|V = v)$ satisfies the scalar density assumption*
4. *$\tilde{P}$ is safe for $\tilde{F}(U|V)|V$ and for each $v \in \text{RANGE}(V)$, $\tilde{P}(U|V = v)$ satisfies the scalar density assumption*
5. *$\tilde{P}$ is pivotally safe for $U|V$ with pivot $\tilde{F}(U|V)$, which is continuous and simple.*

The theorem shows that, whenever pivots are simple (as is often the case), confidence distributions $\tilde{P}$ as defined by (20) are also confidence-safe. If a pivot is nonsimple however, confidence distributions can still be defined via (20) but they may not be confidence-safe under our current definition. An example of such a case is given by the statistical scenario where $\mathcal{M}$ is the 1-dimensional normal family, but the parameter of interest is $U := |\theta|$ rather than θ. As shown by Schweder and Hjort (2016), the confidence distribution $\tilde{P}(U \mid \hat{\theta})$ defined by (20) then gives a point mass $\tilde{P}(U = 0 \mid \hat{\theta} = v) = p > 0$ to $U = |\theta| = 0$ whose size p depends on v. This happens because $\tilde{F}(U \mid v)$ ranges, for such v, not from 0 to 1 but from some $a > 0$ (depending on v) to 1. Then pivotal safety cannot be achieved for $\tilde{P}(U|V)$, since there must be v_1, v_2 with $\tilde{P}(U|V = v_1) \neq \tilde{P}(U|V = v_2)$ whence the definition is not satisfied. Now such confidence distributions based on nonsimple pivots are still useful, and indeed we can prove a weaker form of pivotal and confidence safety for such cases, by replacing safety for $U' \mid [V]$ in Definition 8 by safety for $U' \mid [\![V]\!]$; we will not discuss details here however.

The Hierarchy To see how pivotal and confidence safety fit into the hierarchy of Figure 1, note that Theorem 2 establishes the double arrow between pivotal safety and confidence safety under the scalar density assumption (SDA) — the requirement that $(U', V) \rightsquigarrow U$ in the figure amounts to f_v being a bijection, as we require. The theorem also establishes the relation between calibration and pivotal safety, under the assumption that $\tilde{P}(V)$ has full support and the SDA holds for U. Then the simplest form of calibration, safety for $U \mid [V]$, clearly implies pivotal safety for $U \mid V$ — just take $U' = U$, which is immediately checked to be a pivot. This result trivially extends to the general case of safety for $U \mid [V], V'$ with $V' \neq \mathbf{0}$, this implying pivotal safety with multipivot $U \mid V'$ — we omit the details.

It remains to establish the rightmost column of Figure 1; we will only do this in an informal manner. Schweder and Hjort (2002) (and, implicitly, Hampel (2006)) already note that if $\tilde{P}$ is a confidence distribution for RV U given data V, then it remains a confidence distribution for monotonic functions U' of U, but not for general functions of U. In our framework this translates to, under the scalar density assumption of Section 3.1, that pivotal safety of $U \mid V$ implies pivotal safety for $U' \mid V$ if U' is a 1-to-1 continuous function of U, which readily follows from Definition 7 and Theorem 2 (Definition 7 implies an analogous statement for the discrete case as well). Similarly, it is a straightforward consequence from the definitions that calibration for $U \mid V$ implies calibration for $U' \mid V$, for every U' with $U \rightsquigarrow U'$, not necessarily 1-to-1; yet for U' with $(U, V) \rightsquigarrow U'$, calibration may not be preserved: take e.g. the setting of Example 1 (dilation) with $U' = |V - U|$. Then $\tilde{P}(U' = 1 \mid V = 0) = 0.9$, $\tilde{P}(U' = 1 \mid V = 1) = 0.1$, yet $\mathcal{P}^*$ contains a distribution with $P(U = V) = 1$ and for this P, $P(U' = 1 \mid V) \equiv 0$. If $\tilde{P}$ is valid for $U \mid V$ however, validity is preserved even for every U' with $(U, V) \rightsquigarrow U'$.

3.3 Pivotal Safety and Decisions

Now we consider pivotal safety for RVs U with countable $\text{RANGE}(U)$. The Monty Hall example below shows that in this case, pivotal safety is still a meaningful concept. We first provide an analogue of Theorem 2, in which 'confidence safety' is replaced by something that one might call 'local' confidence safety: safety for a RV U' that determines the *probability of the actually realized outcome* U. To this end, we introduce some notation: for distribution $P \in \mathcal{P}^*$ and RV W, let $p_{[W]}(\cdot)$ be the RV that denotes the probability mass function of W, i.e. $p_{[W]}(w) = P(W = w)$; similarly $\tilde{p}_{[W]}(w) := \tilde{P}(W = w)$. The notation is extended to conditional mass functions as $\tilde{p}_{[U|V]}(u|v) := \tilde{P}(U = u \mid V = v)$. The subscript $[W]$ and $[U|V]$ indicates the RVs under consideration; we will omit them if they are clear from the context. We can think of these mass functions as RVs: for all $z \in \mathcal{Z}$, $p_{[W]}(W)(z) = P(W = W(z))$; $\tilde{p}_{[U|V]}(U|V)(z) = \tilde{P}(U = U(z) \mid V = V(z))$. The difference between RV $\tilde{P}(U \mid V)$ and $\tilde{p}_{[U|V]}(U|V)$ is that the former maps z to the *distribution* $\tilde{P}(U \mid V = V(z))$; the latter maps z to the *probability of a single outcome* $\tilde{P}(U = U(z) \mid V = V(z))$.

Theorem 3 *Let $\mathcal{Z}$ be countable, U be an RV and V a generalized RV. Suppose that for all $v \in \text{RANGE}(V)$, all $p \in [0, 1]$, $\#\{u : \tilde{P}(U = u \mid V = v) = p\} \leq 1$ (i.e. there are no two outcomes to which $\tilde{P}(U \mid V = v)$ assigns the same nonzero probability). Then the following statements are all equivalent:*

1. *$\tilde{P}$ is safe for $\tilde{p}(U|V) \mid [V]$.*
2. *$\tilde{P}$ is pivotally safe for $U \mid V$, with simple pivot $U' = \tilde{p}(U|V)$.*
3. *$\tilde{P}$ is pivotally safe for $U \mid V$ for some simple pivot U''.*

This result establishes that, in the discrete case, if there is *some* simple pivot U'', then $\tilde{p}(U|V)$ is also a simple pivot — thus $\tilde{p}(U|V)$ has some generic status. Compare this to Theorem 2 which established that $\tilde{F}(U|V)$ is a generic pivot in the continuous case.

We now illustrate this result, showing that, for a wide range of loss functions, pivotal safety implies that DM has the right idea of how good her action will be if she bases her action on the belief that $\tilde{P}$ is true — even if $\tilde{P}$ is false. Consider an RV U with countable

range $\mathcal{U} := \text{RANGE}(U)$. Without loss of generality let $\mathcal{U} = \{1, \ldots, k\}$ for some $k > 0$ or $\mathcal{U} = \mathbb{N}$. Let $L : \mathcal{U} \times \mathcal{A} \to \mathbb{R} \cup \{\infty\}$ be a *loss function* that maps outcome $u \in \mathcal{U}$ and action or decision $a \in \mathcal{A}$ to associated loss $L(U, a)$. We will assume that $\mathcal{A} \subset \Delta(\mathcal{U})$ is isomorphic to a subset of the set of probability mass functions on $\mathcal{U}$, thus an action a can be represented by its (possibly infinite-dimensional) mass vector $a = (a(1), a(2), \ldots)$. Thus, L could be any scoring rule as considered in Bayesian statistics (then $\mathcal{A} = \Delta(\mathcal{U})$), but it could also be 0/1-loss, where $\mathcal{A}$ is the set of point masses (vectors with 1 in a single component) on $\mathcal{U}$, and $L(u, a) = 0$ if $a(u) = 1$, $L(u, a) = 1$ otherwise. For any bijection $f : \mathcal{U} \to \mathcal{U}$ we define its extension $f : \mathcal{A} \to \mathcal{A}$ on $\mathcal{A}$ such that we have, for all $u \in \mathcal{U}$, all $a \in \mathcal{A}$, with $a' = f(a)$ and $u' = f(u)$, $a'(u') = a(u)$. Thus any f applied to u permutes this outcome to another u', and f applied to a probability vector permutes the vector entries accordingly.

We say that L is *symmetric* if for all bijections f, all $u \in \mathcal{U}, a \in \mathcal{A}$, $L(u, a) = L(f(u), f(a))$. This requirement says that the loss is invariant under any permutation of the outcome and associated permutation of the action; this holds for important loss functions such as the logarithmic and Brier score and the randomized 0/1-loss, and many others.

We will also require that for all distributions P for U, there exists at least one Bayes action $a_P \in \mathcal{A}$ with $E_P[L(U, a_P)] = \min_{a \in \mathcal{A}} E_P[L(U, a)]$ — which again holds for the aforementioned loss functions. If there is more than one such act we take a_P to be some arbitrary but fixed function that maps each P to associated Bayes act a. In the theorem below we abbreviate $a_{\tilde{P}(U|V)}$ (the optimal action according to $\tilde{P}$ given V, i.e. a generalized RV that is a function of V) to $\tilde{a}_V$.

Theorem 4 *Let $\tilde{P}(U \mid V)$ be a pragmatic distribution where $\mathcal{Z}$ is countable. Suppose that $\tilde{P}(U \mid V)$ is pivotally safe with a simple pivot. Let $L : \mathcal{U} \times \mathcal{A} \to \mathbb{R} \cup \{\infty\}$ be a symmetric loss function as above, and let $\tilde{a}_V$ be defined as above. Then $\tilde{P}$ is safe for $L(U, \tilde{a}_V) \mid [V]$, i.e. for all $v \in \text{RANGE}(V)$, all $P \in \mathcal{P}^*$,*

$$\mathrm{E}_{(U,V)\sim P}[L(U, \tilde{a}_V)] = \mathrm{E}_{\tilde{P}}[L(U, \tilde{a}_V) \mid V = v].$$

Example 11 [Use of Pivots beyond Statistical Inference: Monty Hall] To illustrate Theorem 4, consider again the Monty Hall problem (Example 4) where the contestant chooses door 1. We model this using RV $U \in \{1, 2, 3\}$ representing the door with the car behind it, and $V \in \{2, 3\}$ the door opened by the quiz master; $\mathcal{Z} = \{(1, 2), (1, 3), (2, 3), (3, 2)\}$ and for $z = (u, v)$, $U(z) = u$ and $V(z) = v$ (in this representation it is impossible for the quiz master to open a door with a car behind it). $\mathcal{P}^*$ is the set of all distributions P on $\mathcal{Z}$ with uniform marginal $P(U)$, i.e. as usual, we assume the distribution of the car location to be uniform. Let $\tilde{P}$ be the conditional distribution for $U \mid V$ defined by $\tilde{P}(U = 1 \mid V = 2) = \tilde{P}(U = 1 \mid V = 3) = 1/3$. This distribution can be arrived at using Bayes' theorem, starting with a particular $P \in \mathcal{P}^*$, namely the P with $P(V = 2 \mid U = 1) = P(V = 3 \mid U = 1) = 1/2$, meaning that when the car is actually behind door 1 and the quiz master has a free choice what door to open, he will flip a fair coin to decide. As this game was actually played on TV, it was in fact unknown whether the quiz master actually determined his strategy this way — a quiz master who would want to be helpful to the contestant would certainly do it differently, for example choosing door 3 whenever that is an option. Nevertheless, most analyses, including Vos Savant's original one, assume this particular $\tilde{P}$, and wars have been raging on the wikipedia talk pages as to whether this assumption is justified or not (Gill, 2011).

Interestingly, if we adopt this fair-coin $\tilde{P}$ then $U' = \mathbf{1}_{U=1}$ becomes a discrete simple pivot, in our sense, and $\tilde{P}$ becomes pivotally safe, as is easily checked from Definition 7 and Definition 8: f_v in the definition is given by $f_2(1) = 1, f_2(3) = 0, f_3(1) = 1, f_3(2) = 0$ ($f_2(2)$ and $f_3(3)$ are undefined). Thus $\tilde{P}$ is pivotally safe for Monty Hall and thus Theorem 4 can be applied, showing that, if DM takes decisions that are optimal according to $\tilde{P}$, then these decisions will be exactly as good as she expects them to be for symmetric losses such as 0/1-loss (as in the original Monty Hall problem) but also for the Brier and logarithmic loss. Relatedly, van Ommen et al. (2016) shows that basing decisions on $\tilde{P}$ will lead to admissible and minimax optimal decisions for *every* symmetric loss function (and, in a sequential gambling-with-reinvestment context, even when payoffs are asymmetric). This points to a general link between safety and minimax optimality, which we will explore in future work. Thus, while a strict subjective Bayesian would be *forced* to adopt a single distribution here — for which we do not see very compelling grounds — one can just adopt the uniform $\tilde{P}$ for entirely pragmatic reasons: it will be minimax optimal and as good as one would expect it to be if it were true, even if it's in fact wrong — it may, perhaps, be the case, that people have inarticulate intuitions in this direction and therefore insist that $\tilde{P}$ is 'right'.

4 Beyond Conditioning; Beyond Random Variables

We can think of our pragmatic $\tilde{P}(U|V)$ as probability updating rules, mapping observations $V = v$ to distributions on U. We required these to be compatible with conditional distributions: $\tilde{P}(U|V)$ must always be the conditional of *some* distribution $\tilde{P}$ on $\mathcal{Z}$, even though this distribution may not be in $\mathcal{P}^*$. Perhaps this is too restrictive, and we might want to consider more general probability update rules. Below we indicate how to do this — and present Proposition 3 which seems to suggest that rules that are incompatible with conditional probability are not likely to be very useful. We then continue to extend our approach to update distributions given *events* rather than RVs, leading to the 'sanity check' we announced in the introduction. For simplicity, in this section we restrict ourselves again to V with countable range.

Definition 9 [Probability Update Rule] *Let U be an RV and V be a generalized RV on $\mathcal{Z}$ where* $\text{RANGE}(V)$ *is countable. A* probability update rule $\tilde{Q}(U\|V)$ *is a function from* $\text{RANGE}(V)$ *to the set of distributions on* $\text{RANGE}(U)$*. We call* $\tilde{Q}(U\|V)$ logically coherent *if, for each $v \in \text{RANGE}(V)$, the corresponding distribution on* $\text{RANGE}(U)$*, denoted* $\tilde{Q}(U\|V = v)$*, satisfies*

$$\tilde{Q}(U \in \{u : (u, v) \in \text{RANGE}((U, V))\}\|V = v) = 1. \tag{22}$$

We call $\tilde{Q}(U\|V)$ compatible with conditional probability *if there exists a distribution P on $\mathcal{Z}$ with full support for V (*$\text{SUPP}_P(V) = \text{RANGE}(V)$*) such that* $\tilde{Q}(U\|V) \equiv P(U \mid V)$.

Logical coherence is a weak requirement: if RVs U and V are *logically separated*, i.e. $\text{RANGE}((U, V)) = \text{RANGE}(U) \times \text{RANGE}(V)$ (as is the case in all examples in this paper except Example 11) then clearly every, arbitrary function from $\text{RANGE}(V)$ to the set of distributions on $\text{RANGE}(U)$ is a logically coherent probability update rule. However, if $\text{RANGE}((U, V)) \neq \text{RANGE}(U) \times \text{RANGE}(V)$, then there are logical constraints between U and V. For example, we may have $\mathcal{Z} = \{1, 2\}$, and $U(z) = V(z) = z$ (so that U and V are identical). Then a probability update rule $\tilde{Q}$ with $\tilde{Q}(U = 1\|V = 2) = 1$ would be logically incoherent. Every rule that is

compatible with conditional probability is logically coherent; there does not seem much use in using logically incoherent rules.

For given $\tilde{Q}(U\|V)$ we can now define safety for $U|V$, $U|\langle V\rangle$ and calibration for $U|V$ as before, using Definition 1, 2 and 5. Note however that notions like safety for $U \mid [V]$ and pivotal safety are not defined, since these are defined in terms of marginal distributions for U (or U', respectively), and the marginal $\tilde{Q}(U)$ is undefined for probability update rules Q.

Proposition 3 *If $\tilde{Q}(U\|V)$ is not compatible with conditional probability, then it is not safe for $U \mid \langle V\rangle$ (and hence, as follows directly from the hierarchy of Figure 1, also unsafe for $U \mid V$, and also not calibrated for $U \mid V$).*

Proof: Follows directly from the characterization of safety for $U \mid \langle V\rangle$ given in Proposition 1. □

This result suggests that rules that are incompatible with conditional probability are not likely to be very useful for inference about U; the result says nothing about the weaker notions of safety with $\langle U\rangle$ rather than U on the left though, or with $[\![V]\!]$ instead of $\langle V\rangle$ or V on the left.

Conditioning based on Events Suppose we are given a finite or countable set of outcomes $\mathcal{U}$ with a distribution P_0 on it, as well as a set $\mathcal{V}$ of nonempty subsets of U. We are given the information that the outcome is contained in the set v for some $v \in \mathcal{V}$, and we want to update our distribution P_0 to some new distribution $P_0'(\cdot\|v)$, taking the information in v into account. A lot of people would resort to *naive conditioning* here (Grünwald and Halpern, 2003), i.e. follow the definition of conditional probability and set P_0' to $P_{\text{naive}}(\{u\} \mid v) := P_0(\{u\})/P_0(v)$. We want to see whether such a P_{naive} is *safe*. To this end, we must translate the setting to our set-up: to make a probability update rule in our sense well-defined (Definition 9), we must have a space $\mathcal{Z}$ on which the RV U, denoting the outcome in $\mathcal{U}$, and V, denoting the observed set v, are both defined. To this end we call *any* set $\mathcal{Z}$ such that, for all $u \in \mathcal{U}, v \in \mathcal{V}$ with $u \in v$, there exists a $z \in \mathcal{Z}$ with $U(z) = u$ and $V(z) = v$, a set *underlying $\mathcal{U}$ and $\mathcal{V}$* (we could take $\mathcal{Z} = \mathcal{U} \times \mathcal{V}$, but other choices are possible as well). We then set $\mathcal{P}^*$ to be the set of all distributions P on $\mathcal{Z}$ with marginal distribution P_0 on U and, for all $v \in \mathcal{V}$, $P(U \in v \mid V = v) = 1$. We may now ask whether the naive update,

$$\tilde{Q}(U = u\|V = v) := P_{\text{naive}}(\{u\} \mid v) \tag{23}$$

is safe. The following proposition shows that in general it is not:

Proposition 4 [Grünwald and Halpern (2003), rephrased] *For given P_0, $\mathcal{U}$ and $\mathcal{V}$, let $\mathcal{Z}$ be any set underlying $\mathcal{U}$ and $\mathcal{V}$ and let $\mathcal{P}^*$ be the associated set of distributions compatible with P_0. We have: $\tilde{Q}(U\|V)$ defined as in (23) is the conditional of some distribution $\tilde{Q}$ on $\mathcal{Z}$ that is safe for $U \mid V$ if and only if $\mathcal{V}$ is a partition of $\mathcal{U}$.*

If $\mathcal{V}$ is not a partition of $\mathcal{U}$, then in some cases $\tilde{Q}(U\|V)$ is still compatible with conditional probability; then it is still potentially safe for $U|V$; in other cases it is not even compatible with conditional probability and hence by Proposition 3 guaranteed to be unsafe. The main result of Gill and Grünwald (2008) can be re-interpreted as giving a precise characterization of when this guaranteed unsafety holds.

To illustrate, consider Monty Hall, Example 4 again. In terms of events, $\mathcal{U} = \{1, 2, 3\}$ and $\mathcal{V} = \{\{1, 2\}, \{1, 3\}\}$: if Monty opens door $x, x \in \{2, 3\}$, then the event 'car behind door 1 or x', i.e. $\{1, x\}$ is observed, so $P_{\text{naive}}(\{1\} \mid \{1, 2\}) = P_{\text{naive}}(\{1\} \mid \{1, 3\}) = 1/2$, leading to the common false conclusion that the car is equally likely to be behind each of the remaining closed doors. Clearly though, $\mathcal{V}$ is *not* a partition of $\mathcal{U}$, since it has overlap, so by the proposition, P_{naive} is *unsafe* for $U \mid V$. Intuitively, it is easy to see why: if $U = 1$, the quiz master has a choice what element of $\mathcal{V}$ to present, and may do this by flipping a coin with bias θ. Therefore the set $\mathcal{P}^*$ has an element P_θ corresponding to each $\theta \in [0, 1]$, and the correct conditional distribution $P_\theta(U = 1 \mid V = v)$ depends on θ in a crucial way (and will in fact not be equal to $\tilde{Q}$, no matter the value of θ). But DM need not be concerned with any of these details: what matters is that naive conditioning is not safe, which, by Proposition 3, is immediate from the fact that $\mathcal{V}$ is not a partition of $\mathcal{U}$.

The fact that conditioning is problematic if one conditions on something not equal to a partition has in fact been known for a long time, see e.g. Shafer (1985) for the first landmark reference. Our point is simply to show that the issue fits in well with the safety concept. There is an obvious analogy here with the Borel-Kolmogorov paradox (Schweder and Hjort, 1996) which presumably could also be recast in terms of safety. As Kolmogorov (1933) writes, " The concept of a conditional probability with regard to an isolated hypothesis whose probability equals 0 is inadmissible." Safe probability suggests something more radical: standard conditional probabilities with regard to an isolated hypothesis (event) are *never* admissible — if one does not know whether the alternatives form a partition, setting $\tilde{P}$ to be the standard conditional distribution is inherently unsafe.

5 Parallel, Earlier and Future Work; Open Problems and Conlusion

5.1 Parallel Work: Safe Testing

There is one application of safe probability that has so many implications that we decided to devote a separate paper to it, which we hope to finish soon. This is the use of safety concepts in *testing*, already alluded to in the introduction. Here, let us just very briefly outline some main ideas. Consider a testing problem where we observe data X^N and h_0 stands for a null hypothesis which says that data are a sample from some P_0 belong to a statistical model $\mathcal{M}_0$. For simplicity we will only consider the case of a point null hypotheses in this mini-overview, so $\mathcal{M}_0 = \{P_0\}$ is a singleton. h_1 represents another set of distributions $\mathcal{M}_1$, which may, however, be exceedingly large — in fact it may impossible for us to state it exactly, for it may be, for example, as broad as 'the data are a sample of text in some human language unknown to us'. We associate h_0 with distribution P_0 and corresponding density or mass function p_0, and h_1 with some *single* distribution P_1 with associated p_1. If $\mathcal{M}_1$ is a parametric model, or a large but still precisely specifiable model such as a nonparametric model, then we might take P_1 to be the Bayes marginal distribution under some prior Π, $p_1(X^n) := \int p(X^n) d\Pi(p)$, but other choices are possible as well, and may sometimes even be preferable.

We now define a 'posterior' $\tilde{P}(H \mid X^N)$ by setting

$$\tilde{P}(H = h_0 \mid X^N) := \frac{p_0(X^n)}{p_0(X^n) + p_1(X^n)}, \tag{24}$$

which would coincide with a standard Bayesian posterior based on prior $(1/2, 1/2)$ if we used a p_1 set in the Bayesian way described above. In that special case it also broadly corresponds to the method introduced by Berger et al. (1994) that, in its culminated form (Berger, 2003) provides a testing method that has a valid interpretation within the three major testing-schools: Bayes-Jeffreys, Neyman-Pearson and Fisher. Readers familiar with the MDL (minimum description length) paradigm (Grünwald, 2007) will notice that for every complete lossless code for $X_1, X_2, \ldots$ that encodes X^N with $L(X^n)$ bits, setting $p_1(X^n) = 2^{-L(X^x)}$ provides a probability mass function on sequences of length n. Thus, if one has a code available which one thinks might compress data well, one can set p_1 in this non-Bayesian way. The log-posterior odds $\log \tilde{P}(h_0 \mid X^N)/\tilde{P}(h_0 \mid X^N) = -\log p_1(X^n) + \log p_0(X^n)$ then have an interpretation as the *number of bits saved by compressing the data with the code L compared to the code that would be optimal under P_0*; thus, the approach of Ryabko and Monarev (2005) neatly fits into this framework; so does the Martingale testing approach of Vovk (1993) in which p_1 is determined by a sequential gambling strategy; for any gambling strategy g, there is a corresponding probability mass function p_1 such that the inverse of the (pseudo-) 'Bayes factor'

$$\frac{\tilde{P}(h_0 \mid X^N)}{\tilde{P}(h_0 \mid X^N)} = \frac{p_0(X^n)}{p_1(X^n)} \tag{25}$$

can be interpreted as the amount of money gained by gambling strategy g under pay-offs that would be fair (yield no gain in expectation) if the null P_0 were true: $p_1(X^n)/p_0(X^n)$ is the factor by which one's initial capital is multiplied if one gambles according to g under odds that are fair under h_0, so that the more money gained, the larger the evidence against h_0. For an example of useful non-Bayesian gambling strategies (or equivalently, distributions p_1) we refer to the *switch distribution* of van Erven et al. (2007).

Where Safety Comes In One can now base inferences on $\tilde{P}(H \mid X^N)$ just as a Bayesian would — with the essential stipulation that one only does this for the subset of inferences that once considers *safe* in the appropriate sense. For example, suppose that the data compressor `gzip` compresses our sequence of data substantially more than our null hypothesis P_0, which says that the outcomes are i.i.d. Bernoulli(1/2). Thus, $\tilde{P}(H = h_0 \mid X^N)$ will be exceedingly small, yet the p_1 corresponding to `gzip` may certainly not be considered 'true'. We thus do not want to take the predictions made by p_1 for future data $X_{N+1}, \ldots$ too seriously. We can accomplish this by declaring that p_1 is *not* safe for $X_{N+1} \mid X^N$ relative to $\mathcal{P}^* \mid H = h_1$. Note that we can declare such unsafety without actually precisely specifying $\mathcal{P}^* | H = h_1$, which may be too complicated to do. On the other hand, if we do believe that p_1 accurately describes our knowledge of $\mathcal{M}_1$, e.g. $\mathcal{M}_1$ is small and p_1 is a Bayes marginal distribution based on substantial prior knowledge codified into Π, then we can declare p_1 to be safe relative to $\mathcal{P}^* \mid H = h_1$. We thus have a single framework that encompasses both the Fisherian (falsificationist) and the Bayes/Neyman-Pearson testing paradigms, depending on what inferences we consider safe. On a technical level however, this framework avoids many difficulties of the standard implementations of the Fisherian and the Neyman-Pearson paradigms. Compared to Fisher, we avoid the use of p-values (although the 'Bayes factor' (25) can be interpreted as a *robustified, sample-plan independent* p-value (Shafer et al., 2011, van der Pas and Grünwald, 2014)). We consider this a very good thing in the light of the many difficulties surrounding p-values such as (to mention just two) their dependence on the sampling plan, making them impossible to use in many simple situations and their inter-

pretation difficulties (Berger, 2003, Wagenmakers, 2007). Compared to Neyman-Pearson's original formalism, we do not just get an 'accept' or 'reject' decision, but also a measure of evidence (the (pseudo-) 'Bayes factor') that can be used to infer stronger conclusions as we get stronger evidence — in contrast to conclusions based on p-values, such conclusions often remain *safe* in the appropriate sense.

Indeed, in the second part of this work we consider safety of $\tilde{P}(H \mid X^N)$ in terms of loss functions $L(H, \delta(X^N))$ where $H \in \{h_0, h_1\}$ and $\delta(X^N)$ is the Bayes act based on $\tilde{P}(H \mid X^n)$. In the simplest case takes δ takes values in the decision set $\mathcal{A} = \{\texttt{accept}, \texttt{reject}\}$. We find that, under some conditions on p_1, $\tilde{P}(H \mid X^n)$ is safe for $L(H, \delta(X^N)) \mid [\![X^N]\!]$, i.e. we have safety in the weakest (but still useful) sense defined in this paper. While standard Type-I and Type-II error guarantees of the Neyman-Pearson approach can be recast in this way, safety continues to hold if $\mathcal{A}$ has more than two elements with different losses associated — a realistic situation which cannot be handled by either a Neyman-Pearson or a Fisherian approach. In this situation, making the right decision means one has to take the strength of evidence into account — if there is more evidence against h_0, then the best action to take will have lower loss under h_1 but higher loss under h_0. As soon as there are more than two actions, measuring evidence in terms of unmodified p-values leads to unsafe inferences; yet inferences based on the (pseudo)-posterior tend to remain safe.

Furthermore, we can also check whether we retain safety under *optional stopping* (Berger et al., 1994, Shafer et al., 2011, van der Pas and Grünwald, 2014). We find that, under further conditions on p_0 and p_1, $\tilde{P}$ remains safe for $L(H, \delta(X^N)) \mid [\![X^N]\!]$, even though N is now a RV (determined by the potentially unknown stopping rule) with an unknwon distribution. Interestingly, things get even better with a slight modification of $\tilde{P}$, where we set $\tilde{P}(H = h_0 \mid X^N)$ to $\max\{1, p_0(X^N)/p_1(X^N)\}$, i.e. we use the posterior *odds as* the posterior *probability*. With this 'posterior' we automatically get (weak) safety under arbitrary optional stopping and for essentially arbitrary loss functions – no more conditions on p_0 and p_1 are needed. The reason is that with this choice $\tilde{P}(H \mid X^n)$ becomes bounded by a test martingale in the sense of Vovk (1993) and Shafer et al. (2011). If we want to use the standard posterior as Berger (2003) does, we either need to change the action δ a little (introducing a so-called 'no-decision region', as also done by Berger et al. (1994)) or make strong assumptions about p_0 and p_1.

It is often claimed that optional stopping is not a problem for Bayesians, since the Bayesian inferences do not depend on the sampling plan. For objective Bayesian inference, this is incorrect (priors such as Jeffreys' do depend on the sampling plan); for subjective Bayesian inference, this statement is correct only if one really fully believes one's own subjective prior. As soon as one uses a prior partially for convenience reasons — which happens in nearly all practical scenarios — validity of the conclusions under optional stopping is compromised. Safe testing allows one to establish validity under optional stopping, in a 'weak safety' sense, even in such cases — essentially, one's conclusions will be safe under optional stopping under any P in the set $\mathcal{P}^*$ of possible distributions, not just the single distribution one adopts as a subjective Bayesian.

5.2 Earlier Work and Future Work

The idea that fiducial or confidence distributions can be used for valid assessment of some, not all, RVs or events that can be defined on a domain has been stressed by several authors, e.g. Schweder and Hjort (2002), Xie and Singh (2013), Hampel (2006). The novelty here

is that we formalize the idea and place it in broader context and hierarchy. The idea of replacing sets of distributions by a single representative also underlies the MDL Principle (Rooij and Grünwald, 2011), yet again, without broader context or hierarchy. It is also the core of the pignistic transformation advocated by Smets (1989) as part of his *transferable belief model*, which, apart from the transformation idea, seems to be almost totally different from safe probability however — it would be interesting to sort out the relations though. I already noted in the introduction that my own earlier work contains various definitions of partial notions of safety, but unifying concepts, let alone a hierarchy, were lacking.

Future Work I: Safety vs. Optimality There is one crucial issue though that we neglected in this paper, and that was brought up earlier, to some extent, by (Grünwald, 2000) and (Grünwald and Halpern, 2004): the fact that mere safety isn't enough — we want our pragmatic $\tilde{P}$ to also have optimality properties (see e.g. Example 2 for the trivial weather forecaster who is calibrated (safe) without being optimal). As indicated by Example 11 in the present paper, and also by (Grünwald, 2000) and more implicitly by van Ommen et al. (2016), there is a link between safety and minimax optimality which sometimes can be exploited, but much remains to be done here — this is our main goal for future work.

Future Work II: Objective Bayes Safe Probability may also be fruitfully be applied to objective Bayesian inference. For example, consider inference of a Bernoulli parameter based on an 'objective' Jeffreys' prior $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$. Use of such a prior may certainly be defensible because of its geometric and information-theoretic properties (Rooij and Grünwald, 2011), but what if we have a very small sample of just 1 or even 0 outcomes? Then Jeffreys' prior would tell us, for example, that a bias θ between 0 and 0.01 is 10 times as likely than a bias between 0.495 and 0.505. Most objective Bayesians would probably not be prepared to gamble on that proposition.[9] This is fine, but then what propositions would an objective Bayesian be prepared to gamble on, and what not? Bayesian inference has no tools to deal with this question — and— in a manner similar to characterization of safety for fiducial distributions — safe probability may offer them.

Future Work III: Epistemic Probability More generally, both objective Bayesian and fiducial methods have been proposed as candidates for *epistemic probability* (Keynes, 1921, Carnap, 1950, Hampel, 2006) but it is unclear how exactly such a notion of probability should be connected to decision theory — while a Bayesian or frequentist probability of 0.01 on outcome A implies that a (not too risk-averse) DM would be willing to pay one dollar for a lottery ticket that pays off 200 dollar if A turns out to be the case, for epistemic probability this is not so clear. Safe probability suggests that it might be fruitful to view epistemic probabilities as assuming a willingness to bet on a *strict subset* of all events $\mathcal{A}$ that can be defined on the given space.

Open Problems Other future work involves open problems, as mentioned in the caption of Figure 1. Of particular interest is whether we can extend confidence safety to multidimensional U. Earlier work (Dawid and Stone, 1982, Seidenfeld, 1992) suggests that then in

[9] One might object that an actual value of θ may not even exist, and certainly will never be observed, so one cannot gamble on it. But I could propose this gamble instead: I will toss the biased coin 10000 times, and only reveal to you the final relative frequency of heads. How much would you bet on it being ≤ 0.01?

general, there will be multiple, different choices for $\tilde{P}$, none of which is inherently 'best'. A major additional goal for future work is to identify subjective considerations that may lead one to prefer one choice over another, cf. the idea of 'luckiness' (Rooij and Grünwald, 2011). Another intriguing question is whether safety can be re- construed as an *extension* of measure theory — which has also been designed to restrict the notion of (probability) measures so that they cannot just be applied to any set one likes. Yet another avenue is to extend the definition of pragmatic distributions using upper- and lower expectations, replacing $\tilde{P}$ by a set of distributions $\tilde{\mathcal{P}}$ (this is briefly detailed in Appendix A.1). Then both $\tilde{\mathcal{P}}$ and $\mathcal{P}^*$ would fall into the 'imprecise probability' paradigm; we could still get nontrivial predictions as long as $\tilde{\mathcal{P}}$ is more 'specific' than $\mathcal{P}^*$. Such an extension would hopefully allow us to represent the random-set approach to fiducial inference from Dempster (1968) and its modern extensions, such as the inferential models of Martin and Liu (2013), as an extension of pivotal safety. Here confidence-safe probabilities would be replaced by confidence-safe probability intervals; perhaps one could even arrive at a general description of what applications of Dempster-Shafer theory (Shafer, 1976, Dempster, 1968) are safe at all, and if so, to what degree they are safe.

Acknowledgements This manuscript has benefited a lot from various discussions over the last fifteen years with Philip Dawid, Joe Halpern and Teddy Seidenfeld. Special thanks go to Teddy as well as to Gert de Cooman and Nils Hjort for providing encouragement that was essential to get this work done. This research was supported by the Netherlands Organization for Scientific Research (NWO) VICI Project Nr. 639.073.04.

References

Thomas Augustin, Frank PA Coolen, Gert de Cooman, and Matthias CM Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

O. E. Barndorff-Nielsen and D. R. Cox. *Inference and asymptotics*. Chapman and Hall, 1994.

J. Berger. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1):1–12, 2003.

J.O. Berger, L.D. Brown, and R.L. Wolpert. A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *Annals of Statistics*, 22(4):1787–1807, 1994.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

G.E.P. Box. Robustness in the strategy of scientific model building. In R.L. Launer and G.N. Wilkinson, editors, *Robustness in Statistics*, New York, 1979. Academic Press.

Rudolph Carnap. *Logical Foundations of Probability*. University of Chicago Press, 1950.

David R Cox. Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, 29(2):357–372, 1958.

A Philip Dawid and Mervyn Stone. The functional-model basis of fiducial inference. *The Annals of Statistics*, pages 1054–1067, 1982.

A.P. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77:605–611, 1982. Discussion: pages 611–613.

Arthur P Dempster. A generalization of Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 205–247, 1968.

B. Efron. R. A. Fisher in the 21st century: invited paper presented at the 1996 R. A. Fisher lecture. *Statistical Science*, 13(2):95–122, 1996.

R. A. Fisher. The fiducial argument in statistical inference. *Annals of Eugenics*, 6:391–398, 1935.

R.A. Fisher. Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26: 528–535, 1930.

Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

D.A.S. Fraser. *The structure of inference*, volume 23. Wiley New York, 1968.

D.A.S. Fraser. *Inference and linear models*. McGraw-Hill New York, 1979.

Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of mathematical economics*, 18(2):141–153, 1989.

R.D. Gill and P.D. Grünwald. An algorithmic and a geometric characterization of coarsening at random. *The Annals of Statistics*, 36(5):2409–2422, 2008.

Richard D Gill. The Monty Hall problem is not a probability puzzle — it's a challenge in mathematical modelling. *Statistica Neerlandica*, 65(1):58–71, 2011.

P. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.

P. Grünwald. Safe probability: restricted conditioning and extended marginalization. In *Proceedings Twelfth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2013)*, volume 7958 of *Lecture Notes in Computer Science*, pages 242–252. Springer, 2013.

P. D. Grünwald. Viewing all models as "probabilistic". In *Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT' 99)*, pages 171–182, 1999.

P. D. Grünwald. Maximum entropy and the glasses you are looking through. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI 2000)*, pages 238–246, San Francisco, 2000. Morgan Kaufmann.

P. D. Grünwald and J. Y. Halpern. Updating probabilities. *Journal of Artificial Intelligence Research*, 19:243–278, 2003.

P. D. Grünwald and J. Y. Halpern. When ignorance is bliss. In *Proceedings of the Twentieth Annual Conference on Uncertainty in Artificial Intelligence (UAI 2004)*, Banff, Canada, July 2004.

P.D. Grünwald and J.Y. Halpern. Making decisions using sets of probabilities: Updating, time consistency, and calibration. *Journal of Artificial Intelligence Research (JAIR)*, 42: 393–426, 2011.

F. Hampel. An outline of a unifying statistical theory. In *ISIPTA*, pages 205–212, 2001.

F. Hampel. The proper fiducial argument. In R. Ahlswede, editor, *Information Transfer and Combinatorics*, LNCS, pages 512–526. Springer Verlag, 2006.

Jan Hannig. On generalized fiducial inference. *Statistica Sinica*, pages 491–544, 2009.

J.M. Keynes. *Treatise on Probability*. Macmillan, London, 1921.

A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer-Verlag, 1933.

Dennis V Lindley. Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 102–107, 1958.

Ryan Martin and Chuanhai Liu. Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, 108(501):301–313, 2013.

Judea Pearl. *Causality*. Cambridge university press, second edition, 2009.

Frank P Ramsey. Truth and probability (1926). *The foundations of mathematics and other logical essays*, pages 156–198, 1931.

S de Rooij and PD Grünwald. Luckiness and regret in minimum description length inference. In Prasanta S Bandyopadhyay and M Forster, editors, *Handbook of the Philosophy of Science*, volume 7. Elsevier, 2011.

B Ya Ryabko and VA Monarev. Using information theory approach to randomness testing. *Journal of Statistical Planning and Inference*, 133(1):95–110, 2005.

T. Schweder and N. Hjort. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, 2016.

Tore Schweder and Nils Lid Hjort. Bayesian synthesis or likelihood synthesis: What does Borel's paradox say? Technical Report 46, International Whaling Commission, 1996.

Tore Schweder and Nils Lid Hjort. Confidence and likelihood. *Scandinavian Journal of Statistics*, 29(2):309–332, 2002.

T. Seidenfeld and L. Wasserman. Dilation for convex sets of probabilities. *The Annals of Statistics*, 21:1139–1154, 1993.

Teddy Seidenfeld. R.A Fisher's fiducial argument and Bayes' theorem. *Statistical Science*, pages 358–368, 1992.

G. Shafer. *A mathematical theory of evidence*. Princeton University Press, 1976.

G. Shafer. Conditional probability. *International Statistical Review*, 53(3):261–277, 1985.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, bayes factors and p-values. *Statistical Science*, pages 84–101, 2011.

Philippe Smets. Constructing the pignistic probability function in a context of uncertainty. In *UAI*, volume 89, pages 29–40, 1989.

T.J. Sweeting. Coverage probability bias, objective bayes and the likelihood principle. *Biometrika*, 88(3):657–675, 2001.

Gunnar Taraldsen and Bo Henry Lindqvist. Fiducial theory and optimal inference. *The Annals of Statistics*, 41(1):323–341, 2013.

S. van der Pas and P. Grünwald. Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in nested model selection. Technical Report 1408.5724, Arxiv, 2014.

T. van Erven, P.D. Grünwald, and S. de Rooij. Catching up faster in bayesian model selection and model averaging. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

T. van Ommen, W.M. Koolen, T.E. Feenstra, and P.D. Grünwald. Updating probability beyond conditioning on a partition. *International Journal of Approximate Reasoning*, 2016. To appear.

Piero Veronese and Eugenio Melilli. Fiducial and confidence distributions for real exponential families. *Scandinavian Journal of Statistics*, 42(2):471–484, 2015.

M. vos Savant. Ask Marilyn. *Parade Magazine*, page 15, 1990. There were also followup articles in *Parade Magazine* on Dec. 2, 1990 (p. 25) and Feb. 17, 1991 (p. 12).

V Vovk, A Gammerman, and G Shafer. *Algorithmic learning in a random world*. Springer, New York, 2005.

V.G. Vovk. A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society, series B*, 55:317–351, 1993. (with discussion).

E.J. Wagenmakers. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review*, 14(5):779–804, 2007.

P. Walley. *Statistical Reasoning with Imprecise Probabilities*, volume 42 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 1991.

D. Williams. *Probability with Martingales*. Cambridge Mathematical Textbooks, 1991.

Min-ge Xie and Kesar Singh. Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review*, 81(1):3–39, 2013.

A Technical Extras and Proofs

A.1 Details for Section 2.1: partially specified $\tilde{P}$

As promised in the main text, here we consider $\tilde{P}$ that are only partially specified. We may think of these again as *sets* of distributions, just as we do for $\mathcal{P}^*$. For example, consider an update rule $\tilde{Q}(U\|V)$ as in Definition 9 that is compatible with conditional probability. Such a $\tilde{Q}$ is a prime example of a *partially specified pragmatic distribution*: it is the conditional of at least one distribution P on $\mathcal{Z}$, but there may (and will) be many more, different P for which it is also the conditional. We may thus associate $\tilde{Q}$ with the (nonempty) set $\tilde{\mathcal{Q}}$ of all such distributions P on $\mathcal{Z}$ with $P(U \mid V) = \tilde{Q}(U\|V)$. Then clearly, for every RV U' on $\mathcal{Z}$ with $(U,V) \rightsquigarrow U'$, all $Q_1, Q_2 \in \tilde{\mathcal{Q}}$, we have $Q_1(U'|V) = Q_1(U'|V)$; thus the distribution of such $U_1|V$ is determined by $\tilde{\mathcal{Q}}$; but for U' not determined by (U,V), there may be $Q_1, Q_2 \in \tilde{\mathcal{Q}}$ with $Q_1(U'|V) \neq Q_2(U'|V)$ and we may have to make assessments about U' given V in terms of lower- and upper-expectation intervals $[\inf_{Q\in\tilde{\mathcal{Q}}} \mathrm{E}_Q[U \mid V], \sup_{Q\in\tilde{\mathcal{Q}}} \mathrm{E}_Q[U \mid V]$.

A more involved calculation shows that for all $Q_1, Q_2 \in \tilde{\mathcal{Q}}$, we have $Q_1(U|V') = Q_2(U|V')$ iff $V \rightsquigarrow V' \rightsquigarrow Q(U\|V)$; a condition that also plays a role in Theorem 1 on calibration. One might thus try to state and prove restricted versions of our results, holding for partially specified $\tilde{Q}$ of this form. In practice though, one also encounters other types of partially specified Q (for example, in regression contexts the function $\mathrm{E}_Q[U|V]$ might be used, but no other aspect of Q is relevant). It might thus be more useful to generalize the whole machinery to arbitrary sets of distributions $\tilde{\mathcal{Q}}$; an additional potential advantage is that this might allow us to determine safety of inferential procedures that output sets of probabilities that are nonsingleton yet avoid dilation, such as the inferential model approach of Martin and Liu (2013) and more generally Dempster-Shafer theory. To get a first idea of how this might work, consider the second part of the basic Definition 1. Here we essentially only have to change $\tilde{P}$ to $\tilde{\mathcal{P}}$; nothing else changes:

Definition 10 *Let $\mathcal{Z}$ be an outcome space and $\mathcal{P}^*$ and $\tilde{\mathcal{P}}$ be sets of distributions on $\mathcal{Z}$ as defined in Section 2.1, let U be an RV and V be a generalized RV on $\mathcal{Z}$. We say that $(\mathcal{P}^*, \tilde{\mathcal{P}})$ is* sharply safe *for $\langle U\rangle \mid \langle V\rangle$ if*

$$\textit{for all } P^* \in \mathcal{P}^*, \tilde{P} \in \tilde{\mathcal{P}}: \ \mathrm{E}_{P^*}[U] = \mathrm{E}_{P^*}[\mathrm{E}_{\tilde{P}}[U|V]]. \tag{26}$$

All other definitions may be changed accordingly. We call the resulting notions 'sharply' safe because it requires, for example, safety for $U \mid V$ to imply that *all* distributions in $\tilde{\mathcal{P}}$ agree on $\tilde{P}(U \mid V)$, i.e. their conditional distributions of $U \mid V$ are the same; one could also define weaker notions in which this is only required for *some* $\tilde{P} \in \tilde{\mathcal{P}}$.

A.2 Details for Section 2.2

Proof of Proposition 1, Let k be such that $\textsc{range}(U) \subset \mathbb{R}^k$.

Part 1. Safety of $\tilde{P}$ for $U \mid \langle V\rangle$ implies that for every vector $\vec{a} \in \mathbb{R}^k$ the RV $U_{\vec{a}} = \mathbf{1}_{U\leq\vec{a}}$ satisfies, for all $P \in \mathcal{P}$,

$$\mathrm{E}_{V\sim P}\mathrm{E}_{U_{\vec{a}}\sim P|V}[U_{\vec{a}}] = \mathrm{E}_{V\sim P}\mathrm{E}_{U_{\vec{a}}\sim \tilde{P}|V}[U_{\vec{a}}],$$

which can be rewritten as

$$\sum_{v\in\textsc{range}(V)} P(V=v)[P(U \leq \vec{a} \mid V=v)] = \sum_{v\in\textsc{range}(V)} P(V=v)[\tilde{P}(U \leq \vec{a} \mid V=v)],$$

which in turn is equivalent to $P(U \leq \vec{a}) = P'(U \leq \vec{a})$ with P' as in the statement of the proposition. This shows that safety for $U \mid \langle V \rangle$ implies (9). Conversely, (9) implies that for any function RV U' with $\text{RANGE}(U') \subset \mathbb{R}^{k'}$ with $U \rightsquigarrow U'$, letting f be the function with $U' = f(U)$, we have for every $P \in \mathcal{P}^*$,

$$\mathrm{E}_{U\sim P}[f(U)] = \mathrm{E}_{U\sim P'}[f(U)] = \mathrm{E}_{V\sim P}\mathrm{E}_{U\sim \tilde{P}|V}[f(U)],$$

which implies that $\tilde{P}$ is safe for $\langle U' \rangle \mid \langle V \rangle$, Since this holds for every U' with $U \rightsquigarrow U'$, safety for $U \mid \langle V \rangle$ follows.

Part 2 is just definition chasing. *Part 3* follows as a special case of Part 4 with W in the role of V and $V \equiv \mathbf{0}$. The if-part of *Part 4* is a straightforward consequence of the definition. For the only-if part, note that from the definition of safety for $U \mid [V], W$ we infer that it implies that for all $P \in \mathcal{P}$, for all $v \in \text{SUPP}_{\tilde{P}}(V)$, $w \in \text{SUPP}_P(W)$, for all functions f and RVs $U' = f(U)$,

$$\mathrm{E}_{U\sim P|W=w}[f(U)] = \mathrm{E}_{U\sim \tilde{P}|V=v,W=w}[f(U)]. \tag{27}$$

In particular, this will hold for every $\vec{a} \in \mathbb{R}^k$, for the RV $U_{\vec{a}} = f_{\vec{a}}(U) = \mathbf{1}_{U\leq\vec{a}}$. Then (27) can be written as $P(U \leq \vec{a} \mid W = w) = \tilde{P}(U \leq \vec{a} \mid V = v, W = w)$. Thus, the cumulative distribution functions of $P(U \mid W = w)$ and $\tilde{P}(U \mid V = v, W = w)$ are equal at all $\vec{a} \in \mathbb{R}^k$, so the distributions themselves must also coincide, and (12) follows.

A.3 Details for Section 2.3

Proof of Proposition 2 We let f_0 be the function such that $V \overset{f_0}{\rightsquigarrow}_{\tilde{P}} P(U \mid V)$ and we let $f_1 = f$ be such that $V \overset{f_1}{\rightsquigarrow}_{\tilde{P}} V'$ (f_0 exists by definition, f_1 by assumption). We also let $V'' \equiv \tilde{P}(U \mid V)$ and note that every $v'' \in \text{RANGE}(V'')$ is a probability distribution on U.

We first establish *(1) $\Leftrightarrow$ (2)*. For this, note that since $V \overset{f_1}{\rightsquigarrow}_{\tilde{P}} V'$, we have for all $v \in \text{RANGE}(V)$, for the $v' \in \text{RANGE}(V')$ with $f_1(v) = v'$, that

$$\tilde{P}(U \mid V = v, V' = v') = \tilde{P}(U \mid V = v). \tag{28}$$

If (1) holds, i.e. $\tilde{P}(U \mid V, V')$ ignores V, then the left-hand side in (28) is equal to $\tilde{P}(U \mid V' = v')$, and (2) follows by plugging this into (28). Conversely, if (2) holds, then the right of (28) is equal to $\tilde{P}(U \mid V' = v')$ for all v with $f_1(v) = v'$, and (2) follows by plugging this into (28).

(2) $\Rightarrow$ (3) Suppose that (2) holds. This immediately implies that $V' \overset{f_2}{\rightsquigarrow}_{\tilde{P}} \tilde{P}(U \mid V)$ with $f_2(v') = \tilde{P}(U \mid V' = v')$, which is what we had to prove.

(3) $\Rightarrow$ (4) Suppose that (3) holds. We may thus assume that $V' \overset{f_2}{\rightsquigarrow}_{\tilde{P}} V''(\equiv \tilde{P}(U \mid V))$ for some function f_2. By equivalence (1) $\Leftrightarrow$ (2), which we already proved, it is sufficient to show that for all $v'' \in \text{RANGE}(V'')$, for all $v' \in \text{RANGE}(V')$ with $f_2(v') = v''$, we have $\tilde{P}(U \mid V = v') = \tilde{P}(U \mid V'' = v'')$. Since $\tilde{P}(U \mid V'' = v'') = v''$, it is sufficient to prove that for all $v'' \in \text{RANGE}(V'')$ and for all $v' \in \text{RANGE}(V')$ with $f_2(v') = v''$, we have $\tilde{P}(U \mid V = v') = v''$.

To prove this, fix arbitrary $v'' \in \text{RANGE}(V'')$. For all $v' \in \text{RANGE}(V')$ with $f_2(v') = v''$, for all $v \in \text{RANGE}(V)$ with $f_1(v) = v'$, we must have $f_2(f_1(v)) = v''$ and hence $f_0(v) = v''$, so (by definition of V'') $\tilde{P}(U \mid V = v) = f_0(v) = v''$. Since (from the fact that $V \rightsquigarrow_{\tilde{P}} V'$ and the definition of conditional probability) we can write $\tilde{P}(U \mid V = v') = \sum_{v\in\text{RANGE}(V):f_1(v)=v'} \tilde{P}(U \mid$

$V = v)\alpha_v$ for some weights $\alpha_v \geq 0$, $\sum_{v:f_1(v)=v'} \alpha_v = 1$, and all components of the mixture must be equal to v'', it follows that $\tilde{P}(U \mid V = v') = v''$, which is what we had to prove.

(4) ⇒ (2) We may assume that $V' \overset{f_2}{\rightsquigarrow}_{\tilde{P}} V'' \equiv \tilde{P}(U \mid V)$ for some function f_2, and (by equivalence (1) ⇔ (2) which we already established) that for $v'' \in \text{RANGE}(V'')$, all $v' \in \text{RANGE}(V')$ with $f_2(v') = v''$, $\tilde{P}(U \mid V' = v') = \tilde{P}(U \mid V'' = v'')$. By definition of V'', the latter distribution must itself be equal to v'', so we get:

$$\tilde{P}(U \mid V' = v') = v'', \tag{29}$$

We must also have, for all v with $f_1(v) = v'$, that $f_2(f_1(v)) = v''$, so $f_0(v) = v''$, so, by definition of V'', $\tilde{P}(U \mid V = v) = v''$. Combining this with (29) gives that $\tilde{P}(U \mid V = v) = \tilde{P}(U \mid V' = v')$, and, because $V \rightsquigarrow_{\tilde{P}} V'$, that $\tilde{P}(U \mid V = v, V' = v') = \tilde{P}(U \mid V' = v')$. This must hold for all $v'' \in \text{RANGE}(V'')$, all $v' \in \text{RANGE}(V')$ with $f_2(v') = v''$, and hence simply for all $v' \in \text{RANGE}(V')$ and hence $\tilde{P}(U \mid V, V')$ ignores V.

Final Part By Equivalence (1) ⇔ (2), we have for all $v' \in \text{RANGE}(V')$, all $v \in \text{RANGE}(V)$ with $f_1(v) = v'$, that $\tilde{P}(U \mid V = v) = \tilde{P}(U \mid V' = v')$. Combining this equality with the assumed safety of $\tilde{P}$ for $U \mid V$, we must also have, for all $P \in \mathcal{P}^*$, all $v \in \text{RANGE}(V)$ with $f_1(v) = v'$, that

$$P(U \mid V = v) = \tilde{P}(U \mid V' = v'), \tag{30}$$

But since $P(U \mid V' = v')$ must be a mixture of $P(U \mid V = v)$ over all v with $f_1(v) = v'$ (as in the proof of (3) ⇒ (4) above), and all these mixture components are identical by (30), we get that $P(U \mid V' = v') = \tilde{P}(U \mid V' = v')$. Since this argument is valid for all $v' \in \text{RANGE}(V')$, we have established safety for $U \mid V'$.

Proof of Theorem 1 The result (1) ⇔ (3) is almost immediate: calibration of $\tilde{P}$ is equivalent to having, for each $P \in \mathcal{P}^*$, for each $v'' \in \text{SUPP}_P(V'')$, (note that each such v'' is a probability distribution on U):

$$P(U \mid \tilde{P}(U \mid V_0) = v'') = v'' = \tilde{P}(U \mid V'' = v'').$$

Rewriting the expression on the right of the leftmost conditioning bar as $V'' = v''$, we see that this is equivalent to having

$$P(U \mid V'' = v'') =_P \tilde{P}(U \mid V'' = v'')$$

which by Proposition 1 is equivalent to safety for $U \mid V''$ and so (1) ⇔ (3) follows. From the definition of safety for $U \mid [V], V'$, Definition 1, (3) ⇒ (2) now follows if we can show (by taking, in (2), $V' = V'' = \tilde{P}(U \mid V)$), that (a) $V \rightsquigarrow_{\tilde{P}} V''$ and (b) $\tilde{P}(U \mid V, V'')$ ignores V. The first requirement holds trivially, the second follows from Proposition 2, (3) ⇒ (1), taking again $V' \equiv \tilde{P}(U \mid V)$ (so that automatically $V \rightsquigarrow V'$ and $V' \rightsquigarrow \tilde{P}(U \mid V)$).

It now only remains to show (2) ⇒ (3). So suppose that $\tilde{P}$ is safe for $U \mid V'$ and $\tilde{P}(U \mid V, V')$ ignores V and $V \rightsquigarrow_{\tilde{P}} V'$. By Proposition 2 (1) ⇒ (4), it follows that $\tilde{P}(U \mid V', V'')$ ignores V', where $V'' \equiv \tilde{P}(U \mid V)$. The result now follows by the final part of Proposition 2, applied with V in the proposition set equal to V' and V' in the proposition set equal to V''.

A.4 Details for Section 3

Proof of Theorem 2 *(1) ⇔ (2).* First assume U' is a simple pivot and that pivotal safety holds for U'. Set $f_v(u)$ as in Definition 7 and take it to be increasing for each $v \in \text{RANGE}(V)$ (the decreasing case is analogous). Since U' is a pivot and pivotal safety holds, we have, for all $v \in \text{RANGE}(V)$, $u' \in \text{RANGE}(U')$, $\tilde{F}_{[U'|V]}(u'|v) = F_{[U']}(u')$ so, since f_v is strictly increasing, $\tilde{F}_{[U'|V]}(f_v(u)|v) = F_{[U']}(f_v(u))$ and, because the pivot is simple so f_v is a bijection, $\tilde{F}_{[U|V]}(u|v) = F_{[U']}(f_v(u))$ for all $u \in \text{RANGE}(U \mid V = v)$, so $\tilde{F}_{[U|V]}(u|v)$ is of form (20).

For the converse, assume again that f_v is increasing and take $\tilde{F}_{[U|V]}(u|v)$ of form (20). Then, following the steps above in backward direction, we find that all steps remain valid and show that for all $v \in \text{RANGE}(V)$, $u' \in \text{RANGE}(U')$, $\tilde{F}_{[U'|V]}(u'|v) = F_{[U']}(u')$, which shows that $\tilde{P}$ is pivotally safe for $U|V$ with pivot U'.

(1) ⇒ (4). To show that the SDA (scalar density assumption) is satisfied note that, because U' is a continuous pivot, $P(U')$ satisfies the SDA by definition; because pivotal safety holds, so does $\tilde{P}(U' \mid V = v)$ for each $v \in \text{RANGE}(V)$. Because the pivot U is simple, the function f_v in Definition 7 is a bijection and it follows that $\tilde{P}(' \mid V = v)$ also satisfies SDA for each $v \in \text{RANGE}(V)$.

Now assume that U' is an increasing pivot, i.e. the function $f_v(u) := f(u, v)$ with $U' = f(U, V)$ is increasing in u, for all $v \in \mathcal{V}$ (the decreasing pivot case is proved analogously). For each $b \in [0, 1]$, we have:

$$\begin{aligned}
&\{z \in \mathcal{Z} : \tilde{F}_{[U|V]}(U(z) \mid V(z)) \leq b\}) = \{z \in \mathcal{Z} : \tilde{F}_{[U']}(f(\, U(z), V(z)\,) \mid V(z)) \leq b\} = \\
&\{z \in \mathcal{Z} : \tilde{F}_{[U']}(f(U(z), V(z))) \leq b\} = \{z \in \mathcal{Z} : \tilde{F}_{[U']}(U'(z)) \leq b\} = \\
&\{z \in \mathcal{Z} : F_{[U']}(U'(z)) \leq b\}, \qquad (31)
\end{aligned}$$

where the first equality follows because f_v must be strictly increasing, the second because U' is a pivot, the third is rewriting and the fourth again because U' is a pivot. Because U' is a continuous pivot, it satisfies SDA and thus, for all $P \in \mathcal{P}^*$, $F(U')$, the CDF under P of U', is uniform, so $P(F_{[U']}(U') \leq b) = b$ for all $b \in [0, 1]$. Using (31) now gives that $P(\tilde{F}(U|V) \leq b) = b$.

Since as already established, $\tilde{P}(U \mid V = v)$ satisfies the SDA, we also have $\tilde{P}(\tilde{F}(U|V) \leq b \mid V = v) = b$, for all $v \in \text{RANGE}(V)$. Together these results imply that $\tilde{P}$ is safe for $\tilde{F}(U|V) \mid V$.

(4) ⇒ (5). The third requirement of Definition 7 holds by assumption. To show that the first and second requirements hold, note that by the SDA $f_v(u) := \tilde{F}_{[U|V]}(u|v)$ must be continuous strictly increasing as a function of u on $\text{RANGE}(U)$ for all $v \in \mathcal{V}$, so that $\tilde{F}(U|V)$ is a pivot, and again by the SDA, f_v ranges from 0 to 1 and hence it has an inverse, hence it is a bijection, so that $\tilde{F}(U|V)$ is even a simple pivot.

(5) ⇒ (1) is trivial.

(3) ⇔ (4) Let $U' = \tilde{F}(U|V)$. By Proposition 1, (12), safety for $U' \mid [V]$ is equivalent to having, for all $P \in \mathcal{P}^*$, all $v \in \text{SUPP}_P(V)$, $P(U') = \tilde{P}(U' \mid V = v)$. This in turn is equivalent to having, for all $b \in [0, 1]$, $P(U' \leq b) = \tilde{P}(U' \leq b \mid V = v)$. Since, by the SDA, $\tilde{P}(U' \leq b \mid V = v) = b$, we get that safety for $U' \mid [V]$ is equivalent to having for all $P \in \mathcal{P}^*$, all $v \in \text{SUPP}_P(V)$, all $b \in [0, 1]$, $P(U' \leq b) = b$. But this is just the same as confidence–safety for $U \mid V$ with $a = 0$, which shows (1) ⇒ (2). For the converse, we note that we have just shown that safety for $U' \mid [V]$ implies that for all $P \in \mathcal{P}^*$, all $v \in \text{SUPP}(V)$,

all $0 \leq a < b \leq 1$, that $P(U' \leq b) = b$ and $P(U' \leq a) = a$ whence $P(a < U' \leq b) = b - a$, implying confidence–safety.

Proof of Theorem 3 Let $U' = \tilde{p}(U|V)$ and consider the function f with $U' = f(U, V)$ and let $f_v(u)$ be as in Definition 7. The following fact is immediate by the condition of 'uniqueness of nonzero probabilities' imposed on $\tilde{P}(U \mid V = v)$:

Fact 1. For each $v \in \text{RANGE}(V)$, f_v is an injection.

We then find, by Definition 7 and 8 that U' is a simple pivot and $\tilde{P}$ is pivotally safe iff for all $P \in \mathcal{P}^*$, for all $v \in \text{RANGE}(V)$,

$$P(U') = \tilde{P}(U') = \tilde{P}(U' \mid V = v),$$

which, by Proposition 1, (12), is equivalent to $\tilde{P}$ being safe for $U' \mid [V]$. This establishes (1) $\Leftrightarrow$ (2). The implication (2) $\Rightarrow$ (3) is trivial (take U' as pivot). Thus it only remains to show:

(3) $\Rightarrow$ (1): Let U'' be a simple pivot for $U \mid V$ and suppose that pivotal safety holds with pivot U''. We first show that for each $p \in [0, 1]$, $\tilde{p}_{[U|V]}(U|V) = p \Leftrightarrow \tilde{p}_{[U'']}(U'') = p$, i.e.

$$\{z \in \mathcal{Z} : \tilde{p}_{[U|V]}(U(z)|V(z)) = p\} = \{z \in \mathcal{Z} : \tilde{p}_{[U'']}(U''(z)) = p\}. \tag{32}$$

To see this, note that, because for each $v \in \text{RANGE}(V)$, the mapping $f_v(u) := f(u, v)$ is a bijection from $\text{RANGE}(U \mid V = v)$ to $\text{RANGE}(U'')$, we have

$$\{z \in \mathcal{Z} : \tilde{p}_{[U|V]}(U(z)|V(z)) = p\} = \{z \in \mathcal{Z} : \tilde{p}_{[U''|V]}(f_{V(z)}(U(z))|V(z)) = p\} =$$
$$\{z \in \mathcal{Z} : \tilde{p}_{[U'']}(f_{V(z)}(U(z))) = p\} = \{z \in \mathcal{Z} : \tilde{p}_{[U'']}(U''(z)) = p\},$$

where the first equality follows from Fact 1 above, the second because of pivotal safety, which imposes that $\tilde{P}(U'' \mid V = v) = \tilde{P}(U'')$ for all $v \in \text{RANGE}(V)$, and the third by definition of $f_v(u)$. Thus, (32) follows, and it implies that the two events in (32) must have the same probability under any single probability measure on $\mathcal{Z}$, in particular under $\tilde{P}(\cdot \mid V = v)$ for all $v \in \text{RANGE}(V)$ and for all $P \in \mathcal{P}^*$, i.e.

$$\tilde{P}(\{z \in \mathcal{Z} : \tilde{p}_{[U|V]}(U(z)|V(z)) = p\} \mid V = v) = \tilde{P}(\{z \in \mathcal{Z} : \tilde{p}_{[U'']}(U''(z)) = p\} \mid V = v) \tag{33}$$
$$P(\{z \in \mathcal{Z} : \tilde{p}_{[U|V]}(U(z)|V(z)) = p\}) = P(\{z \in \mathcal{Z} : \tilde{p}_{[U'']}(U''(z)) = p\}). \tag{34}$$

Since U'' is a pivot, $\tilde{P}(U'' \mid V = v)$ is the same for all $v \in \text{RANGE}(V)$ and equal to $\tilde{P}(U'')$ and also to $P(U'')$, for all $P \in \mathcal{P}^*$. Combining this with (33) we find that

$$\tilde{P}(\{z \in \mathcal{Z} : \tilde{p}_{[U|V]}(U(z)|V(z)) = p\} \mid V = v) = P(\{z \in \mathcal{Z} : \tilde{p}_{[U'']}(U''(z)) = p\}).$$

Rewriting this further using (34) gives

$$\tilde{P}(\{z \in \mathcal{Z} : \tilde{p}_{[U|V]}(U(z)|V(z)) = p\} \mid V = v) = P(\{z \in \mathcal{Z} : \tilde{p}_{[U|V]}(U(z)|V(z)) = p\}),$$

i.e., setting $U' = \tilde{p}_{[U|V]}(U|V)$, we find that for all $v \in \text{RANGE}(V)$, $\tilde{P}(U' \mid V = v) = P(U')$; thus $\tilde{P}$ is pivotally safe for $U|V$ with simple pivot U', and *(1).* follows.

Proof of Theorem 4 By assumption there is some simple pivot $U' = f(U, V)$, such that for each $v \in \text{RANGE}(V)$, the function f_v on $\text{RANGE}(U \mid V = v)$ defined as $f_v(u) = f(u, v)$ is a bijection to $\text{RANGE}(U')$. We now fix some function $g : \mathcal{U}' \to \text{RANGE}(U)$ that is 1-to-1 (an injection, not a bijection). Such a function must exist; we can, for example, take $f_{v_0}^{-1}$ for arbitrary but fixed v_0 which exists because f_{v_0} must be a bijection by definition. Also note that for any bijection $f : \mathcal{U} \to \mathcal{U}$ and its extension to $\mathcal{A}$ as defined in the main text, we have, for every distribution P on $\mathcal{U}$ with mass function p and $\tilde{a}_{P(U)}$ denoting the function from $P(U)$ to a Bayes act for $P(U)$ (which we assume to exist), by symmetry of the loss:

$$\sum_{u \in \mathcal{U}} P(f(u)) \cdot L(f(u), f(\tilde{a}_{P(U)})) = \sum_{u \in \mathcal{U}} P(u) \cdot L(u, \tilde{a}_{P(U)}) =$$
$$\min_{a \in \mathcal{A}} \sum_{u \in \mathcal{U}} P(u) \cdot L(u, a) = \min_{a \in \mathcal{A}} \sum_{u \in \mathcal{U}} P(f(u)) \cdot L(f(u), a) = \sum_{u \in \mathcal{U}} P(f(u)) \cdot L(f(u), f(\tilde{a}_{P(f(U))})),$$

hence, combining the leftmost and righmost expression,

$$f(\tilde{a}_{P(U)}) = \tilde{a}_{P(f(U))}. \tag{35}$$

Now repeatedly using symmetry of the loss function and (35), we have:

$$\sum_{u \in \mathcal{U}} \tilde{P}(U = u \mid V = v) \cdot L(u, \tilde{a}_v) = \sum_{u \in \mathcal{U}} \tilde{P}(\{z : U(z) = u\} \mid V = v) \cdot L(u, \tilde{a}_v) =$$
$$\sum_{u \in \mathcal{U}} \tilde{P}(\{z : f_v(U(z)) = f_v(u)\} \mid V = v) \cdot L(g(f_v(u)), g(f_v(\tilde{a}_v))) =$$
$$\sum_{u' \in \mathcal{U}'} \tilde{P}(\{z : U'(z)) = u'\} \mid V = v) \cdot L(g(u'), a_{\tilde{P}(g(f_v(U))|V=v)}) =$$
$$\sum_{u' \in \mathcal{U}'} \tilde{P}(\{z : U'(z)) = u'\}) \cdot L(g(u'), a_{\tilde{P}(g(U')|V=v)}) = \sum_{u' \in \mathcal{U}'} P(\{z : U'(z)) = u'\}) \cdot L(g(u'), a_{\tilde{P}(g(U'))}) =$$
$$\sum_{u' \in \mathcal{U}', v \in \text{RANGE}(V)} P(\{z : U'(z)) = u', V(z) = v\}) \cdot L(g(u'), a_{\tilde{P}(g(U'))}) =$$
$$\sum_{u' \in \mathcal{U}', v \in \text{RANGE}(V)} P(\{z : U'(z)) = u', V(z) = v\}) \cdot L(g(u'), a_{\tilde{P}(g(U')|V=v)}) =$$
$$\sum_{u' \in \mathcal{U}', v \in \text{RANGE}(V)} P(\{z : f_v^{-1}(U'(z)) = f_v^{-1}(u'), V(z) = v\}) \cdot L(g(f_v(f_v^{-1}(u'))), a_{\tilde{P}(g(f_v(U))|V=v)}) =$$

$$\sum_{u \in \mathcal{U}, v \in \text{RANGE}(V)} P(\{z : U(z) = u, V(z) = v\}) \cdot L(g(f_v(u)), a_{\tilde{P}(g(f_v(U))|V=v)}) =$$
$$\sum_{u \in \mathcal{U}, v \in \text{RANGE}(V)} P(\{z : U(z) = u, V(z) = v\}) \cdot L(g(f_v(u)), g(f_v(a_{\tilde{P}(U|V=v)}))) =$$
$$\sum_{u \in \mathcal{U}, v \in \text{RANGE}(V)} P(\{z : U(z) = u, V(z) = v\}) \cdot L(u, \tilde{a}_v) = \text{E}_{U,V \sim P}[L(U, \tilde{a}_V)],$$

and the result follows.